Title
Bayesian Methods for Mortality Estimation

Author
Alexander, Monica

Publication Date
2018

Peer reviewed|Thesis/dissertation

Bayesian Methods for Mortality Estimation

by

Monica Alexander

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Demography

in the

Graduation Division

of the

University of California, Berkeley

Committee in charge:

Professor Joshua R. Goldstein, Chair
Professor Kenneth W. Wachter
Professor Dennis Feehan
Professor Jasjeet S. Sekhon

Spring 2018

Abstract

Bayesian Methods for Mortality Estimation

by

Monica Alexander

Doctor of Philosophy in Demography

University of California, Berkeley

Professor Joshua R. Goldstein, Chair

The development of mortality models is important in order to reconstruct historical processes, understand current patterns and predict future trends. Mortality models are particularly useful when the available data are sparse, unreliable or incomplete. Traditionally, mortality patterns in data-rich populations were used to observe mathematical or empirical regularities, which could be applied to data-sparse populations. However, as a wider variety of data have become available, the focus of model building has shifted to developing flexible models that perform well in a variety of contexts.

This dissertation introduces Bayesian methods of mortality estimation in three contexts where the available data are imperfect. The first paper develops a method to estimate subnational mortality in situations with small populations and highly-variable data. The second paper develops a unified modeling framework to estimate and project neonatal mortality in all countries worldwide, including those with limited and poor-quality data. The third paper introduces a new dataset to study mortality inequalities in the United States, and develops methods to deal with the truncated and censored mortality information that is available. In all three contexts, the modeling approaches combine strengths from traditional demographic models, which capture mortality regularities across age, with the flexibility of Bayesian frameworks, which allow for multiple data sources to be incorporated, information to be shared across time and space, and uncertainty to be assessed.

Acknowledgements

Undertaking a PhD has been an extremely challenging, but enjoyable and rewarding experience. I have been lucky to have had the support and mentorship of a lot of great people along the way.

I would like to thank my committee for their academic guidance. Thanks to Ken Wachter for sharing his wisdom and experience, Dennis Feehan for his suggestions and encouragement, and Jas Sekhon for his perspective and support. In particular, thanks to my committee chair, Josh Goldstein, whose mentorship and original way of thinking helped me broaden my skills as a researcher. I really benefited from our discussions, and appreciated his motivation and support of my endeavors throughout the PhD.

In addition to my committee, I was lucky to have the support of other researchers, who gave me opportunities along the way: Magali Barbieri, Robert Chung, Rayid Ghani and, in particular, Emilio Zagheni and Leontine Alkema. Emilio's encouragement over the past few years has been invaluable, and I really enjoyed and learnt a lot from working together. Meeting Leontine in the Spring of 2014 was when I discovered my research area of interest and since then, her support, encouragement, and the opportunities she has given me have been incredible. Emilio and Leontine exemplify what a difference early-career researchers can make to their students, and their mentorship inspires me to try and encourage and support my future students in the same way.

I have had some great classmates and peers during the PhD: the statistics MA group, the DSSG group in Chicago, Emily and Niamh in Amherst, and my fellow demography graduate students. Their intelligence makes it hard not to suffer from imposter syndrome. Thanks especially to the Plan Bs — Paul Chung, Robert Pickett and Leslie Root — for the demography and R discussions, the puns, coffee, board games, beers and burritos. Thanks for making the Berkeley journey such an enjoyable one. See you at BDE in the future.

Thanks to my family for their continued love and support. To Mum and Dad, who gave me a breadth of opportunities to learn from an early age, always encouraged me to work hard and to stick at things, and have always supported me, no matter what. Finally, to Rohan, who knew I could do a PhD even before I realized it was an option. Thanks for listening to research ideas and practice talks, reading and editing papers, for the road trips, ramen and BBQ, and for always believing in me.

Contents

Chapter 1

Introduction

A core focus of demography as a discipline is building models to describe aggregate population processes using a reduced set of parameters. Models can be used to describe historical processes, decompose current patterns, and to forecast. Mortality research is particularly suited to the development of models to explain underlying processes, as mortality patterns tend to exhibit relatively regular shapes across age and change gradually over time.

Mortality models are particularly useful when the available data are sparse, unreliable or incomplete. Traditionally, mortality patterns in data-rich populations were used to observe mathematical or empirical regularities, which could be applied to data-sparse populations. However, as a wider variety of data have become available, the focus of model building has shifted to developing flexible models that perform well in a variety of contexts. As the demand for accurate and timely estimates increases, there is a need to develop models which incorporate both traditional demographic methods and more flexible statistical frameworks, to overcome known data issues.

The papers in this dissertation introduce methods of mortality estimation in three contexts where the available data are imperfect. The first paper develops a method to estimate subnational mortality in situations with small populations and highly variable data. The second paper develops a unified modeling framework to estimate and project neonatal mortality in all countries worldwide, including those with limited and poor-quality data. The third paper introduces a new dataset to study mortality inequalities in the United States, and develops methods to deal with the truncated and censored mortality information that is available. In all three contexts, the modeling approaches combine strengths from traditional demographic models, which capture mortality regularities across age, with the flexibility of Bayesian frameworks that allow for multiple data sources to be incorporated, information to be shared across time and space, and uncertainty to be assessed.

The following sections motivate the development of new methods of mortality es-

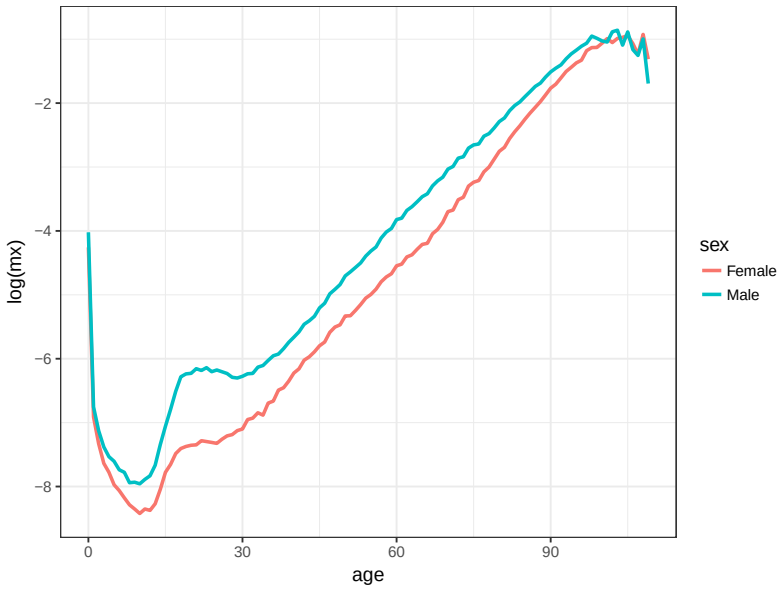


Figure 1.1: Log mortality rates, United States 1975. Data are from the Human Mortality Database.

timation. Traditional approaches to mortality estimation using both mathematical and empirical models are briefly discussed. The need for incorporating uncertainty in mortality estimation, and the increased use of Bayesian methods, is then highlighted. Finally, the three areas of research included in this dissertation are summarized.

1.1 Estimating the mortality curve

Mortality data are usually available in the form of the number of deaths in a particular age group and the corresponding population at risk (usually approximated by the mid-year population). As such, the focus of estimation is often the mortality rate (i.e. deaths per population at risk per time). In human populations, mortality rate curves by age have a characteristic 'J' shape when plotted on the log scale: relatively high in the first year of life, decreasing, then increasing across age from about age 10. At the oldest ages, there is often an observed deceleration of mortality. For example, Fig. 1.1 shows the mortality rates by age and sex for the United States in 1975.

Some important goals of mortality modeling include describing the shape of these mortality curves, estimating and projecting mortality patterns over time, and investigating differences in mortality patterns across different populations. Often we are interested in estimating mortality patterns in all three of these dimensions: age, time and space. As shown in Fig. 1.1, the shape of the mortality curve by age is inherently non-linear and so not necessarily straightforward to fit with a reduced set of parameters. Models to estimate mortality can broadly be categorized as either

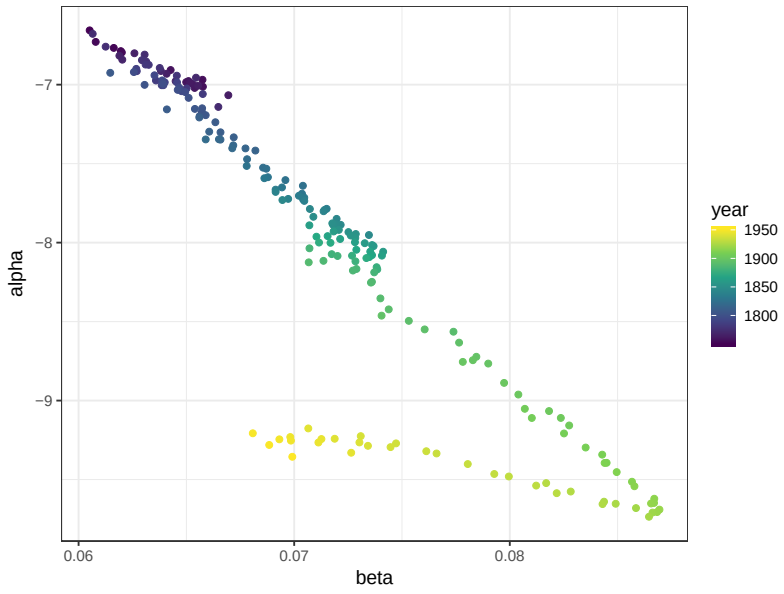


Figure 1.2: Gompertz parameters for Sweden, 1750-1950 cohorts. Data are from the Human Mortality Database. The model is fit to mortality above age 30.

mathematical or empirical.

1.1.1 Mathematical models

One approach is to estimate mortality curves based on a functional or mathematical model that is governed by a set of parameters. It is generally the case that each parameter has a demographic interpretation and so looking at differences in these parameters across time, space or subpopulations gives a summary measure of mortality differences. The most well-known and simplest of all mathematical models is the Gompertz model (Gompertz (1825)). Gompertz observed that the pattern of mortality by age, x, follows a near geometric series, and as such age-specific mortality rates $m(x)$ can be neatly approximated by two parameters

$$m(x) = \alpha e^{\beta x}$$

where α represents a baseline mortality level and β represents the rate of mortality increase over age. Trends in estimates of α and β can be explored and projected to infer future mortality levels. For example, Fig. 1.2 shows estimated Gompertz parameters α and β over cohorts for 1750-1950 in Sweden. There is a clear relationship between the parameter estimates through time, with a decrease in α and increase in β over cohorts, up until around 1900. Since then, α has increased slightly but β has started to decrease. These trends are consistent with a decrease and compression in mortality over time, followed by a slow down of compression in more recent cohorts. Given the regularity in trends over time, these parameters could easily be projected to infer future trends.

On the log scale, the Gompertz model corresponds to linearly increasing mortality over age. Referring to Fig. 1.1, this assumption is reasonable for adult ages (say from around age 30 to 95). However, the fit is poor at relatively young and old ages. There are many other mathematical models that are more complex than the Gompertz model and aim to better capture mortality trends at younger and older ages. For example, the Gompertz-Makeham model includes an additional parameter that is age-independent and aims to capture background/extrinsic mortality (Makeham (1860)). The Siler model (Siler (1983)) is a five parameter model that includes components for infant, adult and senescent mortality. The Heligman-Pollard model (Heligman and Pollard (1980)) includes eight parameters, which includes additional flexibility to account for the younger-adult accidental mortality hump.

There are many other models along these lines (see Preston et al. (2000); Wachter (2014); Feehan (2017) for reviews). While mathematical models of mortality are useful in summarizing and comparing broad mortality trends across populations and time periods, it is generally difficult to capture all the variation in the highly non-linear mortality curve with a tractable functional form. While the flexibility of fit increases as more parameters are included, this also makes estimation more difficult, and parameter estimates can be unstable due to high correlation between parameters (Missov et al. (2015)).

1.1.2 Empirical models

Another approach to estimating mortality is to model age patterns based on empirical regularities in observed data, rather than a theoretical functional form. The idea behind these sorts of approaches is that age patterns in mortality, while inherently non-linear, exhibit strong similarities across different populations.

The most common empirical method of mortality estimation is using model lifetables. For this approach, patterns observed in high-quality data are used as a basis for producing estimates of mortality in populations where data are sparse or poor quality. Model lifetables are produced as a standard mortality table by age which can then be applied to other populations based on partial mortality patterns that are observed.

Ansley Coale and Paul Demeny published a set of model life tables in 1966 (Coale et al. (1966)). They identified different patterns of mortality based on four distinct regions: North, South, East, and West. Each had a characteristic pattern of child mortality and a different relationship between child mortality and adult mortality.

The original Coale-Demeny lifetables were based almost exclusively on European data. They were derived from a set of 192 life tables, including several time periods (39 from before 1900 and 69 from after the Second World War) and mostly from Western countries. Europe, North America, Australia and New Zealand contributed

a total of 176 tables. The United Nations (United Nations (1982)) published a model lifetable system using data across a broader range of countries, including developing countries. More recently, authors such as Wilmoth et al. (2012) and Clark (2016) have used data available through the Human Mortality Database (HMD (2018)) to build more flexible systems of life tables.

Another empirical approach to mortality estimation is relational models. In particular, Brass (1971) introduces a relational logit model whereby survivorship by age l_x is modeled as a function of a mortality standard, Y_x, and two parameters, α and β:

$$\text{logit } l_x = \alpha + \beta Y_x.$$

Brass proposed a standard, Y_x, however in practice any standard can be used. The parameters α and β have demographic interpretations, affecting the overall level of survival and the relative importance of infant and old-age mortality, respectively. In a similar way to the Gompertz model, Brass parameters can be estimated over time, used to summarize differences across populations and potentially forecast mortality trends. The idea of relational mortality models has been extended in ways such as including four parameters (Ewbank et al. (1983)) and modeling old-age mortality (Himes et al. (1994)).

In general, empirical mortality models are useful because they allow non-linear trends across age to be expressed with relatively few parameters. The empirical approach exploits the fact that age patterns of mortality exhibit characteristic shapes across different subpopulations and these shapes can be applied to situations where only limited data are available. However, empirical models are relatively rigid in terms of being able to follow the data that do exist for hard-to-measure populations. The classical model lifetable set-up does not allow for 'new' patterns of mortality by age, and assumes countries with poor-quality data follow the experience of countries that have good data. Estimates are highly sensitive to the choice of standard, and recent mortality trends (such as the HIV/AIDS epidemic) cannot be taken into account unless that information is contained in the standard.

There are other methods of mortality estimation which provide more flexibility in fit. In particular, there has been an increased use of penalized splines, or P-splines models to fit mortality across age, time and space (Currie et al. (2004)). P-splines regression methods produce non-linear smooth fits in mortality patterns using basis-splines as covariates in a regression, with the smoothness of the resulting fit penalized (Eilers and Marx (1996); Currie and Durban (2002)). Recent research utilizes the flexibility of P-splines to fit and smooth mortality patterns in a wide range of contexts (Eilers et al. (2008); Ugarte et al. (2010); Ouellette and Bourbeau (2011); Camarda (2012)). The main drawback of using splines models is that there is no restriction on the age pattern of mortality, and so there is the potential for demographically implausible age schedules, especially in contexts where death is a rare outcome, the populations are small or the available data are sparse.

1.2 Incorporating uncertainty

An important consideration that is largely omitted from traditional methods of mortality estimation is the level of uncertainty, or error, around the estimates and projections produced.

There are different types of uncertainty that need to be taken into account. Firstly, there is the potential for error in the observed data. For survey data, observations may have both sampling error, based on the smaller sample size compared to the population, and non-sampling error, if the survey is not representative. For deaths that are recorded in vital registration systems, even if there is complete registration of deaths, there is still stochastic error, i.e. the observed number of deaths may be higher or lower than expected, purely by chance. This is particularly the case when deaths are a rare outcome or populations are small. There is also model-based uncertainty, that is, uncertainty in the model parameters fit to the data. For example, in Fig. 1.2, if the Gompertz model was a better fit for some cohorts than others, the standard errors around the α and β estimates would be lower. In addition, projecting mortality trends into the future, methods should allow for a certain degree of uncertainty to be propagated through time.

In general, traditional methods of mortality estimation do not lend themselves particularly well to account for different types of uncertainty. Often demographic modeling is performed in a purely deterministic sense, with no incorporation or allowance for any uncertainty. In addition, if the available data are of poor quality and errors are high, such as the case in many developing countries, observations are often dismissed and estimates from model lifetables are believed instead. However, producing and reporting estimates of uncertainty is important in demographic research because it communicates a measure of the level of confidence we have about trends in the past and what may happen in the future.

There has been an increasing move towards including stochastic methods to fully capture uncertainty in estimates and projections. In terms of modeling mortality, many estimation methods are now performed in a Poisson framework to account for the stochasticity in deaths (e.g. Congdon et al. (1997); Foreman et al. (2012); Camarda (2012)). The work of Lee and Carter (1992) in forecasting US mortality rates using time series methods provided a new standard in stochastic demographic forecasting, with many authors extending this method in more recent work (see for example Booth et al. (2006); Girosi and King (2007); Delwarde et al. (2007); Lee (2015); Wiśniowski et al. (2015)).

1.2.1 Bayesian modeling

The increase in probabilistic modeling in mortality and demographic estimation has been coupled with an increase in the use of Bayesian methods. In classical frequentist

methods, the focus is on the likelihood function of data y given parameters θ: the goal of estimation is then to estimate the parameters θ which best explain the observed data y. These parameters are treated as fixed, but unknown. In contrast, Bayesian approaches allow the underlying parameters of interest θ to be expressed as random variables. This means that we can have prior belief about the probability distribution of the possible values of θ.

In Bayesian statistics, we still observe data y and want to estimate parameters θ that best explain the observed data, i.e. $\theta|y$. Bayes rule states the posterior distribution of the parameters, $P(\theta|y)$ is

$$P(\theta|y) = \frac{P(y|\theta) \cdot P(\theta)}{P(y)}$$

where $P(y|\theta)$ is the likelihood function, $P(\theta)$ is the prior distribution on the parameters of interest and $P(y)$ is the marginal probability of the data.

Although Bayesian probability is more than 250 years old, the use of Bayesian methods for estimation did not increase rapidly until the 1990s, largely as a consequence of increased computing power. In the context of demographic estimation, Bayesian methods were first used as a method for producing population projections with uncertainty (Daponte et al. (1997); Alho and Spencer (2006)). More recently, there has been a rapid increase in Bayesian demography, due in part to this method being used to produce probabilistic projections of populations for countries worldwide by the United Nations Population Division. This method of population projection, developed by Adrian Raftery and colleagues (Alkema et al. (2011); Raftery et al. (2012); Raftery et al. (2013); Azose et al. (2016)) models fertility rates, life expectancy and migration in one probabilistic framework. Other researchers have increasingly used Bayesian techniques for the estimation and projection of all demographic components, including mortality (Girosi and King (2008); Alkema and New (2014); Wiśniowski et al. (2015)), fertility (Schmertmann et al. (2013); Schmertmann et al. (2014)) and migration (Bijak (2010)). See Bijak and Bryant (2016) for a good review.

There are several advantages of using Bayesian methods for demographic modeling. Firstly, many different sources of uncertainty can be incorporated into the modeling framework and projected through time. Parameter uncertainty is reflected through prior distributions. The stochasticity of the process, as well as additional sampling and non-sampling errors can be captured as part of the likelihood function (or 'data model'). In this way, data from multiple sources can be incorporated into one framework, and the different quality of each source can be taken into account.

Another important aspect of the Bayesian approach is that it allows prior information to be incorporated into the model. The central idea of Bayesian analysis is that the resulting posterior estimates of the parameters θ are some combination of the information in the prior and what is observed in the data. The more data available, the more parameter estimates will be influenced by the observed patterns. In

situations where there are limited data available, the estimates are more influenced by prior specifications. Thus, the Bayesian approach is particularly useful in sparse data situations, providing a robust way of combining the little information that is available with plausible outcomes based on prior specifications. This is useful for 'pooling' information across different dimensions: age, space and time. For example, countries or areas with fairly limited data available can be partially informed by trends in similar countries or areas. In practice, priors are usually specified in a hierarchical way; that is, the parameters governing the priors are themselves given hyper-priors governed by hyper-parameters. This hierarchical structure is useful in capturing regularities across age, space and time. For example, if mortality was being estimated using a Gompertz model within a Bayesian set-up, the autocorrelation of mortality trends through time could be captured by placing a time series model on the Gompertz parameters α and β.

More broadly, Bayesian analysis in demographic estimation provides a useful framework to extend traditional mortality estimation methods in a more flexible way, combining different sources of information and incorporating different sorts of uncertainty into the one modeling framework. In this way, the strengths of traditional mortality estimation in capturing regularities in mortality are retained, while estimates are better informed by data that are available, while taking uncertainty into account.

1.3 Bayesian methods for mortality estimation

In this dissertation, methods of mortality estimation are developed that utilize a combination of traditional demographic approaches and Bayesian methods in order to produce robust estimates and levels of uncertainty in key indicators over time. The three papers focus on estimation methods for situations where the data available are limited or poor quality for different reasons: small populations; sparse data availability in developing countries; and a large but censored dataset. Each of these three issues is becoming increasingly important in the goal of producing timely and accurate mortality estimates in many contexts.

The first paper develops a model for subnational mortality estimation. National-level mortality indicators often mask the variation in outcomes that can exist within a country. As such, there has been an increasing focus on understanding spatial disparities in mortality outcomes and how these vary across time. Reliable estimates of mortality indicators at the subnational level are essential for monitoring these trends and inequalities. However, one of the difficulties in producing such estimates is the presence of small populations, particularly when death rates are relatively low. In these contexts, the stochastic variation in death counts is relatively high, and thus the underlying mortality levels are unclear. By just considering raw death rates, areas with smaller populations are more likely to look higher or lower than average, purely by chance.

In the first paper, a flexible modeling framework to estimate subnational age-specific mortality rates is presented. The model builds on demographic information about characteristic age patterns in mortality curves, which are constructed using principal components from a set of reference mortality curves. Information on mortality rates are pooled across geographic space and are smoothed over time.

The second paper develops a model for estimating neonatal mortality rates in all countries. With the emergence of the Millennium Development Goals (MDGs) and, more recently, the Sustainable Development Goals (SDGs), there has been increased need for timely estimates and projections of key demographic and health indicators in countries worldwide. However, in practice it is often the countries with the poorest countries that have the least data available on key indicators, and the data that are available are often of poor quality. Traditional demographic approaches, for example using model lifetables to produce a single estimate of mortality outcomes, are not sufficient in tracking progress towards achieving global health goals.

A particularly important set of development goal indicators are related to child mortality. The MDG 4 called for a two-thirds reduction in under-five mortality (U5MR) between 1990 and 2015. As the U5MR decreases, the share of neonatal deaths, i.e. deaths occurring in the first month, tend to increase. As such Goal 3 of the SDGs explicitly includes a neonatal target, with the aim to reduce the neonatal mortality rate to at least as low as 12 deaths per 1000 live births in all countries by 2030.

The second paper presents a Bayesian splines regression model for estimating neonatal mortality rates (NMR) for all countries. In the model, the relationship between NMR and U5MR is assessed and used to inform estimates, and spline regression models are used to capture country-specific trends. As such, the resulting NMR estimates incorporate trends in overall child mortality while also capturing data-driven trends.

The final paper introduces new data and methods in order to study mortality inequalities in the United States. It is important to study differences in mortality outcomes by demographic and socioeconomic characteristics in order to fully understand national trends in mortality over time. An issue with studying mortality inequalities, particularly by socioeconomic status (SES), is that there are few micro-level data sources available that link an individual's SES with their eventual date of death. There has been an increasing amount of mortality inequality research that makes use of large-scale administrative datasets, or linking multiple individual datasets to match SES with mortality. The final chapter introduces a new matched dataset, 'CenSoc', which uses the full-count 1940 Census to obtain demographic, socioeconomic and geographic information, linked to the Social Security Deaths Masterfile, to obtain mortality information.

While large in size, and rich in socioeconomic information, observations of deaths in CenSoc are truncated and censored, and so mortality estimation is not necessarily

straightforward. As such, the final paper also develops mortality estimation methods to better use the 'deaths without denominators' information contained in CenSoc. Bayesian hierarchical methods are presented to estimate truncated death distributions over age and cohort. The model frameworks are based on traditional mortality models, such as the Gompertz model and a principal components relational model. The Bayesian set-up allows for additional prior information in mortality trends to be incorporated and estimates of life expectancy and associated uncertainty to be produced.

In summary, this dissertation presents three Bayesian methods of mortality estimation to overcome issues of high stochastic variability, sparse data, and truncated observations. While the methods presented solve estimation issues in specific contexts, the overarching theme is using modeling approaches that utilize the strength of demographic methods in combination with the flexibility of Bayesian methods. Incorporating demographic knowledge into modeling frameworks ensures estimates and projections have plausible patterns across time and particularly across age. The use of Bayesian frameworks allows different sources of information to be combined, data of varying quality to be weighted accordingly, and uncertainty to be better accounted for and propagated through time. The philosophy of integrating demographic and Bayesian modeling approaches, and thinking about how to balance model-driven versus data-driven results, is useful in developing model frameworks in a wide range of contexts.

Chapter 2

A flexible Bayesian model for estimating subnational mortality

2.1 Introduction[1]

In order to effectively study health disparities within a country, it is important to obtain reliable subnational mortality estimates to quantify geographic differences accurately. There is a large demand for estimates of small-area mortality as indicators of overall health and well-being, as well as for natural experiments that exploit policy changes at local levels. Reliable mortality estimates for regional populations could help to better understand how place-of-residence and communities can affect health status, through both compositional and contextual mechanisms (Macintyre et al. (2002)).

One of the difficulties in producing mortality estimates for subnational areas is the presence of small populations where the stochastic variation in death counts is relatively high. For example, 10%, or around 300, of US counties have populations of less than 5,000, and 1% of counties have less than 1,000 people (US Census Bureau (2014)). The resulting mortality rates in small areas are often highly erratic and may have zero death counts, meaning the underlying true mortality schedules are unclear.

The aim of this paper is to formulate a model for estimating mortality rates at the subnational level, across geographic areas with a wide variety of population sizes and death counts. Resulting estimates would be useful for guiding future policy efforts to improve the health of populations and to investigate the historical impact of public health interventions and changes in the structure of local health programs. In this article, we focus on developing the methodology to produce age- and sex-specific mortality rates and the approach is tested on simulated data that mimic US

[1]This chapter has been previously published as Alexander et al. (2017).

counties, and on real data for French *départements*. However, the model is flexible enough to be used in a wide range of situations.

There has been a growing literature in the field of small-area mortality estimation. However, the demand for accurate, reliable and consistent estimates has not yet been met. The traditional life table approach assumes that deaths y_{ax} in a population in area a at age x are poisson distributed $y_{a,x} \sim \text{Poisson}(P_{a,x} \cdot m_{a,x})$ where $P_{a,x}$ is the population at risk and $m_{a,x}$ is the mortality rate. The maximum likelihood estimate of the mortality rate for area a at age x is

$$\hat{m}_{ax} = \frac{y_{a,x}}{P_{a,x}}. \tag{2.1}$$

This approach essentially involves estimating as many fixed-effect parameters m_{ax} as there are data points. In addition, this estimation process makes no reference to or use of the information about mortality rates at other ages, in other areas or at other time points. Confidence intervals can be derived based on the distribution of deaths, but for small populations, stochastic variation is high and so confidence intervals and standard errors are large. Estimating mortality rates in small populations therefore requires different types of approaches.

To avoid issues that arise in small-area mortality estimation, a common approach is to aggregate mortality data across multiple years or across space to form larger geographic areas. For example, recent work by Chetty et al. (2016) and Currie and Schwandt (2016) look at mortality inequalities in the US using deaths and income measures at the county level, but data are either aggregated across space and time (Currie and Schwandt (2016)) or results are not published for smaller populations (in the case of Chetty et al. (2016), where the minimum population size is 20,000). However, given the information lost in aggregating data from smaller areas, there is value in employing other techniques to infer mortality levels and trends.

One option that has been employed is to treat each small population as a stand-alone population and model accordingly using traditional model life table approaches. For example, Bravo and Malta (2010) outline an approach for estimating life tables in small populations, applied to regional areas of Portugal. They estimate Gompertz-Makeham functions via generalized linear models, with an adjustment at older ages. Another approach by Jarner and Kryger (2011) involves estimating old-age mortality in small populations by first estimating parameters of a frailty model using a larger reference population. However, approaches that treat small populations separately do not account for the likely relationships between the regional population estimates or patterns over time. Other approaches in the US have used county-level covariates such as socio-economic status and education to predict county-level life expectancy (Ezzati et al. (2008); Dwyer-Lindgren et al. (2016); Kindig and Cheng (2013); Kulkarni et al. (2011); Srebotnjak et al. (2010)). However, there are issues with using the resulting estimates to infer relationships between health, poverty rates and education without concerns of endogeneity.

In this paper, we propose a new model that relies on a Bayesian hierarchical frame-

work, allowing information on mortality to be shared across time and through space. This helps to inform the mortality patterns in smaller geographic areas, where uncertainty around the data is high. The modeling process produces uncertainty intervals around the mortality estimates, which can then be translated into uncertainty around other life-table quantities (for example, life expectancy). As well as producing uncertainty intervals around the final estimates, the modeling process also involves the estimation of other meaningful variance parameters that may relate to variation in mortality within, or across, states. A Bayesian hierarchical approach has been used in estimating mortality in small populations previously (e.g. see Congdon (2014)). However, our modeling approach is novel in that it combines demographic knowledge about regularities in age-specific mortality with the flexibility of a Bayesian hierarchical structure.

The remainder of the paper is structured as follows. In the next section, the methodology for estimating age-specific mortality rates is described. The model is then applied to two different data situations – simulated and real – and results are discussed. Performance of the model is evaluated through coverage and mean-squared-error measures. The paper concludes with a discussion of model performance and contributions.

2.2 Method

We propose a model that has an underlying functional form that captures regularities in age patterns in mortality. We then build on this functional form within a Bayesian hierarchical framework, penalizing departures from the characteristic shapes across age, as well as sharing information across geographic areas and ensuring a relatively smooth trend in mortality over time.

Bayesian hierarchical models have previously been used in a wide range of demographic applications. For example, Bayesian hierarchical models are used by the United Nations Population Division to produce probabilistic projections of population and life expectancy (Raftery et al. (2012); Raftery et al. (2013)). Alkema and New (2014) developed a Bayesian hierarchical model for estimating the under-five mortality rate for all countries worldwide. There are many other examples in the fields of mortality, fertility and migration (e.g. Alkema et al. (2012); Bijak (2008); Congdon (2009); King and Soneji (2011); Sharrow et al. (2013)). Our approach has similarities to applications by authors in cause of death mortality estimation (Girosi and King (2008)) and cohort fertility projection (Schmertmann et al. (2014)), but with a focus on addressing small-area estimation issues, rather than forecasting.

2.2.1 Model set-up

Let $y_{x,a,t}$ be the deaths at age x in area a at time t. We assume that deaths are Poisson distributed

$$y_{x,a,t} \sim \text{Poisson}(P_{x,a,t} \cdot m_{x,a,t}), \tag{2.2}$$

where $m_{x,a,t}$ is the mortality rate at age x, area a and time t and $P_{x,a,t}$ is the population at risk at age x, area a and time t. We estimate mortality rates at ages 0, 1, 5, and then in 5-year intervals up to 85+.

The $m_{x,a,t}$ are modeled on the log-scale as:

$$\log(m_{x,a,t}) = \beta_{1,a,t} \cdot Y_{1x} + \beta_{2,a,t} \cdot Y_{2x} + \beta_{3,a,t} \cdot Y_{3x} + u_{x,a,t}, \tag{2.3}$$

where Y_{px} is the pth principal component of some set of standard mortality curves, and $u_{x,a,t}$ is a random effect. The use of principal components has similarities with the Lee-Carter approach (Lee and Carter (1992)). Principal components create an underlying structure of the model in which regularities in age patterns of human mortality can be expressed. Many different kinds of shapes of mortality curves can be expressed as a combination of the components. Incorporating more than one principal component allows for greater flexibility in the underlying shape of the mortality age schedule.

Principal components are obtained via a singular value decomposition (SVD) on a matrix which contains a set of standard mortality curves. The SVD is a factorization of a matrix, which is useful in explaining the main aspects of variation of the data. For example, for the application to simulated US counties below, we used US state mortality rates from 1980–2010. Let $\mathbf{X}$ be a $N \times G$ matrix of log-mortality rates, where N is the number of observations and G is the number of age-groups. In the US states case, we had $N = 50 \times 31 = 1550$ observations of $G = 19$ age-groups. The SVD of $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{UDV}', \tag{2.4}$$

where $\mathbf{U}$ is a $N \times N$ matrix, $\mathbf{D}$ is a $N \times G$ matrix and $\mathbf{V}$ is a $G \times G$ matrix. The first three columns of $\mathbf{V}$ (the first three right-singular values of $\mathbf{X}$) are Y_{1x}, Y_{2x} and Y_{3x}.[2]

The first three principal components for US state mortality curves from 1980–2010 are shown in Fig. 2.1 below.[3] They were based on mortality curves on the log scale, three of which are shown in the top lefthand graph: Florida in 1980, Hawaii in 1991 and California in 2010. In total, there are $50 \times 31 = 1550$ of these curves, from which the principal components are derived. Broadly, the first principal component describes the overall mortality curve. The second principal component allows

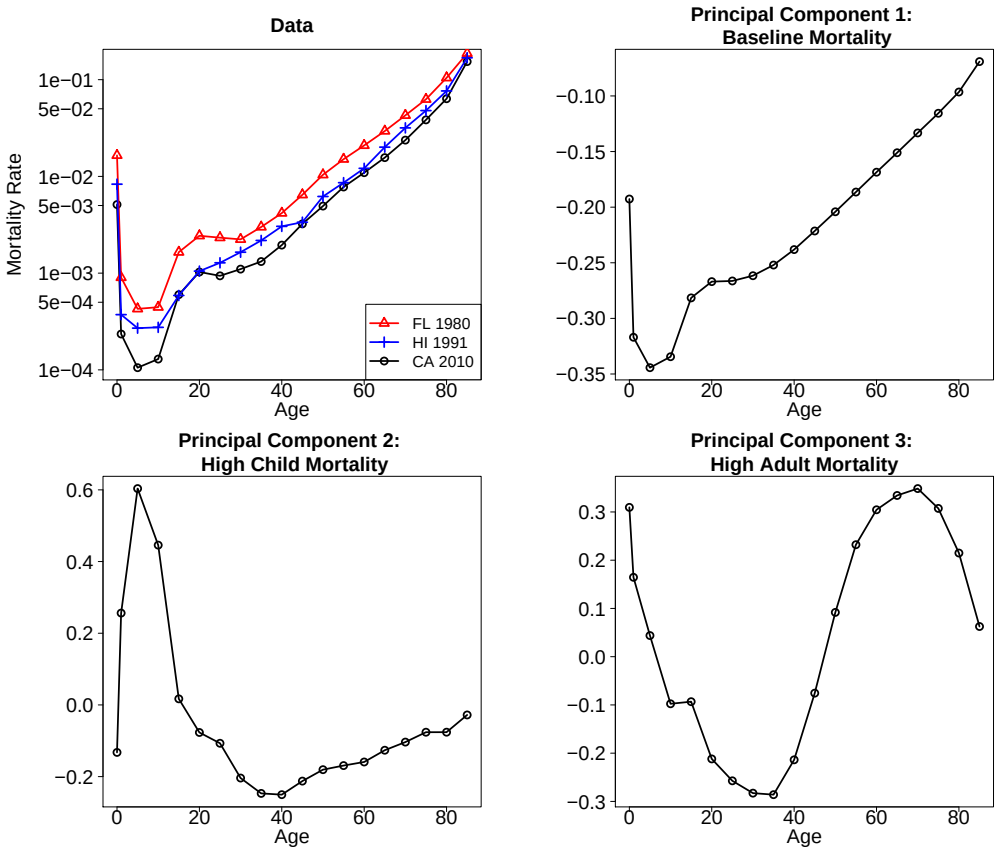
[2]Throughout the paper, we refer to the Y_{px}'s as principal components for simplicity, even though technically Y_{px} is really the $p-$th vector of principal component loadings.

[3]Data on age-specific mortality rates are available through the Centers for Disease Control and Prevention (CDC) Wide-ranging OnLine Data for Epidemiologic Research (WONDER) tool: `https://wonder.cdc.gov/`.

mortality at younger ages to be higher in relation to adult mortality. The third principal component allows adult mortality to be higher in relation to mortality at young and old ages. For example, in some regions of a country, child mortality might be relatively higher than the baseline schedule. In other regions, where prevalence of deaths due to accidents is higher, adult mortality would be higher than the baseline pattern.

The components capture overall patterns of mortality well; a wide range of different mortality curves can be expressed as a linear combination of these three components. It is possible to include more or less than three principal components in the model for $\log(m_{x,a,t})$. The more principal components used, the more flexible the fit. However, after experimenting with a wide range of standard mortality curves, we found that higher-order principal components generally did not display any regular patterns across age but instead picked up on residual variance in the dataset, which has limited use for modeling purposes. Including the first three principal components in the model allows for a reasonably flexible fit while only including components that have some demographic interpretation.

Figure 2.1: Example data and principal components of (logged) US state mortality schedules, Males, 1980–2010.



The addition of the random effect term $u_{x,a,t}$ in the expression for $\log(m_{x,a,t})$ accounts for potential over-dispersion of deaths, i.e. the case where the variance in deaths is greater than the mean, which otherwise would not be expected given the assumption

of Poisson-distributed deaths (Congdon (2009)). It is assumed that these random effects are centered at zero with an associated variance:

$$u_{x,a,t} \sim N(0, \sigma_x^2). \tag{2.5}$$

The variance parameter varies by age group, allowing heterogeneity in some age groups to be greater than in others. In practice it is often the younger age groups, with the lowest levels of mortality, that have the highest variation.

2.2.2 Pooling information across geographic area

To allow for information on the level and shape of mortality to be shared across geographic space, we assume that the coefficients $\beta_{p,a,t}$ for a particular area are drawn from a common distribution centered around a state (or country) mean:

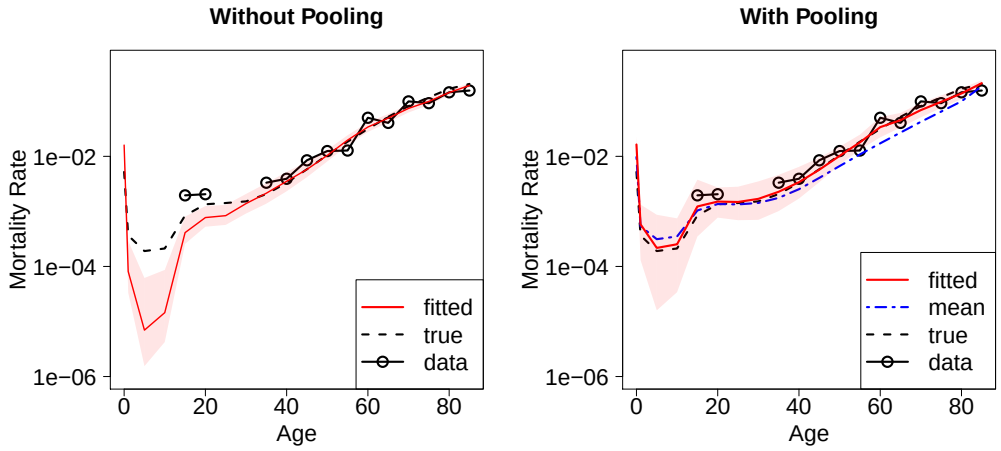
$$\beta_{p,a,t} \sim N(\mu_{\beta_{p,t}}, \sigma^2_{\beta_{p,t}}) \tag{2.6}$$

where p indicates principal component ($p = 1, 2, 3$). Larger areas in terms of population size (and death counts) have a bigger effect on the overall means. The less data available on deaths in an area, the closer the parameter estimates are to the mean parameter value. In this way, mortality patterns in smaller areas are partially informed by mortality patterns in larger areas. At the same time, mortality patterns in larger areas borrow little information from the pooling process and are largely determined by their own observed death counts.

The influence of the geographic pooling is illustrated in Fig. 2.2. The charts illustrate observed, true and fitted log-mortality rates for a hypothetical county with a population of 5,000 males. The dashed line is the true underlying log-mortality rate. The circles represent the observed log-mortality rates; these were simulated from the true rates using Eqn. 2.2. Where there are gaps, the observed death count was zero. The solid line and associated shaded area is the fit and 95% credible intervals. The graph on the left showed a fit without geographic pooling, while the graph on the right shows a fit with geographic pooling. The right-hand graph has an additional 'mean' line. This line was derived from the mean parameters, $\mu_{\beta_{p,t}}$, which are informed by all counties in the state. The effect of pooling is seen most in log-mortality rates at younger ages, where rates are low. As many of the younger age groups have observed zero death counts, the unpooled model estimates log-mortality rates that are much lower than the true rates. The pooled model estimate is 'pulled' up closer to the mean estimate, benefiting from information on young-age mortality from other counties.

In practice, the mean parameter values $\mu_{\beta_{p,t}}$ could be determined from any plausible group of areas which share similar characteristics. We have tested the model using state level mean parameter values, but other options include grouping areas by a smaller location scale, by rural/urban area within state, or by a common age distribution.

Figure 2.2: Illustrating the effect of geographic pooling. Model fitted without (left) and with (right) geographic pooling. Data are simulated assuming a total population of 5,000.



2.2.3 Smoothing across time

We assume the parameters governing the shape of the mortality curve, $\beta_{p,a,t}$, change gradually and in a relatively regular pattern over time. We impose this smoothing by penalizing the second-order differences across time in the mean parameters:

$$\mu_{\beta_{p,t}} \sim N\left(2 \cdot \mu_{\beta_{p,t-1}} - \mu_{\beta_{p,t-2}}, \sigma^2_{\mu_{\beta_{p,t}}}\right) \tag{2.7}$$

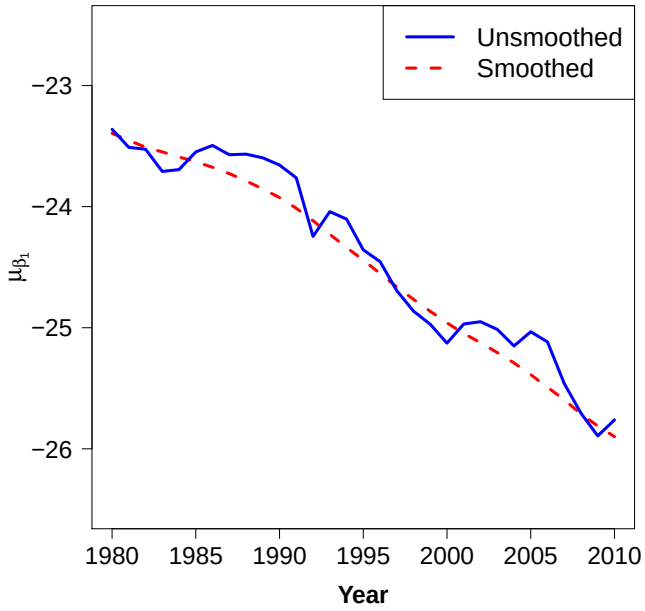
for $p = 1, 2, 3$. This set up is penalizing differences from a linear trend in the mean parameters. Smoothing the mean parameters, rather than the actual parameters $\beta_{p,a,t}$, still allows for mortality trends to depart from a smooth trajectory if suggested by the data. For example, if a particular area suffered from a influenza outbreak thus making mortality higher than in previous years, the $\beta_{p,a,t}$ terms would allow for higher mortality.

The effect of smoothing parameters over time is shown in Fig. 2.3. The graph shows the estimated median value of the parameter $\mu_{\beta_{1,t}}$ over 31 years in a simulated US county model. The solid line shows the estimates without smoothing, while the dashed line shows the effect of smoothing.

2.2.4 Adding constraints to the model for total areas

While mortality rates are estimated for subnational populations, it is important that these mortality rates, when aggregated to the state or national level, are consistent with the mortality rates observed at the aggregate level. To ensure this is the case, we add a constraint to the model which specifies that the number of deaths in a state (or country) is Poisson distributed with a rate equal to the sum of all estimated

Figure 2.3: Illustrating the effect of smoothing over time. The solid line shows median estimates for $\mu_{\beta_{1,t}}$ from a model without smoothing imposed. The dashed line shows median estimates for $\mu_{\beta_{1,t}}$ from a model with smoothing imposed according to Eqn. 2.7.



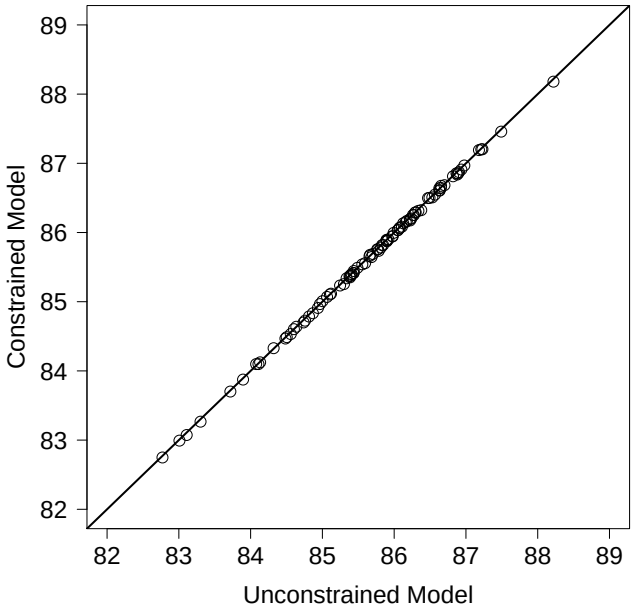
deaths in all areas:

$$\sum_{a=1}^{A} y_{x,a,t} \sim \text{Poisson}\left(\sum_{a=1}^{A}(P_{x,a,t} \cdot m_{x,a,t})\right) \tag{2.8}$$

where A is the total number of areas.

We assume that the deaths $y_{x,a,t}$ are conditionally independent; that is, given values of $P_{x,a,t}$ and $m_{x,a,t}$, the $y_{x,a,t}$'s are independent, and thus the sum is also Poisson distributed. In practice, the added constraint makes little difference to the estimates for any specific area. However, it ensures consistency between regional and national estimates. We believe that this is an important feature of the model that is relevant for most applications. Fig. 2.4 shows life expectancy at birth estimates for females for each French *département* in 2008 (data are described below in section 2.3.2). The resulting estimates show only negligible differences between the constrained and unconstrained models.

Figure 2.4: Female life expectancy at birth estimates for France, 2008 (years): constrained versus unconstrained model.



2.2.5 Priors and Implementation

Non-informative priors were put on variance parameters. Operationally, we used a uniform distribution between 0 and 40 for the standard deviations:

$$\sigma_{\beta_{p,t}} \sim U(0, 40) \tag{2.9}$$
$$\sigma_{\mu_{\beta_{p,t}}} \sim U(0, 40) \tag{2.10}$$
$$\sigma_x \sim U(0, 40). \tag{2.11}$$

The non-informative nature of these prior distributions means posterior estimates of the variance parameters were not constrained in any way. Choosing even more spread-out uniform distributions as the priors (with a maximum larger than 40) had no effect on the final estimates.

The model was fitted in a Bayesian framework using the statistical software R. Samples were taken from the posterior distributions of the parameters via a Markov Chain Monte Carlo (MCMC) algorithm. This was performed using JAGS software Plummer (2003). Standard diagnostic checks using trace plots and the Gelman and Rubin diagnostic (Gelman and Rubin (1992)) were used to check convergence.

2.2.6 Model Summary

The set of equations below summarizes the entire model set-up. Deaths are assumed to be Poisson distributed, and the mortality rates that govern the Poisson process are specified by a model, which is hierarchical in structure. The first level (Eqn. 2.13) gives an expression for $\log(m_{x,a,t})$. The second level (Eqns. 2.14 and 2.15) specifies distributions for the parameters in the first level ($\beta_{p,a,t}$ and $u_{x,a,t}$). Finally, the third level (Eqns. 2.15–2.19) specifies the distribution for the parameters in the second level ($\mu_{\beta_{p,t}}$ and the variance terms).

$$
\begin{aligned}
y_{x,a,t} &\sim \text{Poisson}(P_{x,a,t} \cdot m_{x,a,t}) && (2.12)\\
\log(m_{x,a,t}) &= \beta_{1,a,t} \cdot Y_{1x} + \beta_{2,a,t} \cdot Y_{2x} + \beta_{3,a,t} \cdot Y_{3x} + u_{x,a,t} && (2.13)\\
\beta_{p,a,t} &\sim N(\mu_{\beta_{p,t}}, \sigma^2_{\beta_{p,t}}) && (2.14)\\
\mu_{\beta_{p,t}} &\sim N\left(2 \cdot \mu_{\beta_{p,t-1}} - \mu_{\beta_{p,t-2}}, \sigma^2_{\mu_{\beta_{p,t}}}\right) && (2.15)\\
u_{x,a,t} &\sim N(0, \sigma^2_x) && (2.16)\\
\sigma_{\beta_{p,t}} &\sim U(0, 40) && (2.17)\\
\sigma_{\mu_{\beta_{p,t}}} &\sim U(0, 40) && (2.18)\\
\sigma_x &\sim U(0, 40) && (2.19)
\end{aligned}
$$

2.3 Results

2.3.1 Simulated data

In order to test the model, we created a simulated data set of deaths and populations that mimic counties within US states. The 'true' mortality rate in a county is based on a specified population and age structure, and the mortality rate in the state. The mortality curve for a county can be altered to change shape via a Brass relational model setup, assuming that:

$$\log\left(\frac{l_x}{1 - l_x}\right) = \alpha + \beta \cdot Y_x \qquad (2.20)$$

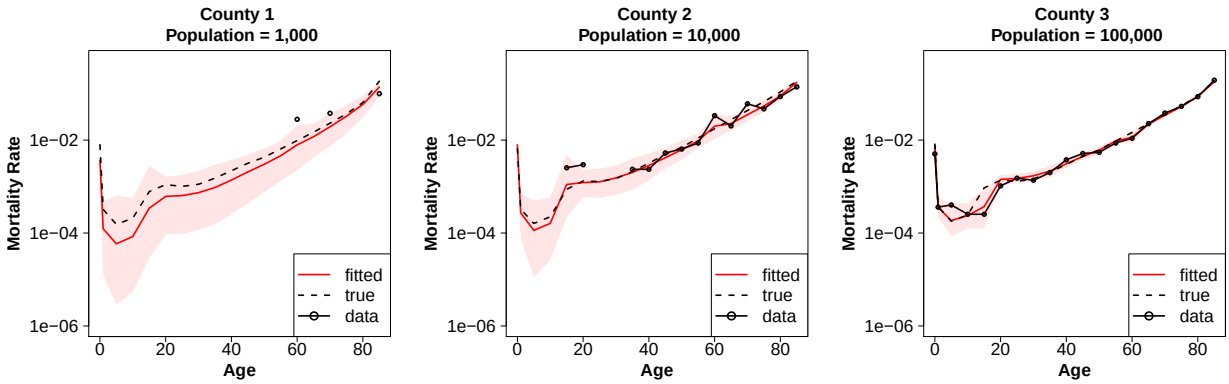
where l_x is the survivorship at age x in the county and Y_x is the survivorship at age x in the state of interest. To alter the shape of the survivorship curve for a particular county, the values of α and β were changed. The values of α and β were chosen randomly from the ranges $[-0.75, 0.75]$ and $[0.7, 1.3]$, respectively. These ranges of α and β values were chosen because they translate to a reasonable range of age-specific mortality curves commonly observed. The survivorship rates were then converted to mortality rates using standard life table relationships.

Once the 'true' mortality rate schedules were obtained, we simulated deaths according to the relationship shown in Eq. 2.2. A range of population sizes were tested,

with the minimum county size being 1,000 people of a particular sex. At this small population size, many simulated death counts for particular age groups are equal to zero.

Fig. 2.5 shows the true, simulated ('observed') data and estimated mortality rates on the log scale in three hypothetical counties within the same state but with different population sizes. The points show the observed data, which is simulated from the true underlying mortality rate, shown by the dashed line. For the smallest county, which has 1,000 people, many of the observed death counts are zero, so the data do not show up on the log scale. The solid line shows the estimated log-mortality rates, and the corresponding shaded area shows the 95% credible intervals. As the size of the county increases, the mortality pattern in the observed data becomes more regular. As such, the uncertainty around the estimates decreases with increased population size.

Figure 2.5: True, simulated and estimated mortality rates for three hypothetical counties.



Evaluation of model performance

In order to evaluate model performance, we compared the model fit to the fit of a simple Loess smoother and Brass model. The Loess approach does not incorporate any pooling of information or demographic regularities across age. The Brass estimation process uses Eqn. 2.20. The methods of estimation were compared using the simulated data, where the true value of the mortality rates was known. For each area and time, we estimated the root mean squared error (RMSE), defined as

$$\text{RMSE} = \sqrt{\frac{1}{G}\sum_{x=1}^{G}(\hat{m}_x - m_x^*)^2}, \tag{2.21}$$

where $\hat{m}_x$ is the estimated mortality rate at age x, m_x^* is the true mortality rate and G is the number of age groups.

Table 2.1 below shows the average RMSE for the three fits to a simulated dataset containing 60 counties over 31 years (12 counties per size group). In all cases, the

RMSE decreases as county size increases. This is intuitive because, as the county population increases, there are fewer zero death counts and so more information about the shape of the mortality curve. The average RMSE for the model is always lower than Loess or Brass, irrespective of county size. Although the Brass RMSE seems reasonable, it is most likely because the data were generated using a Brass relational model.

Table 2.1: Average Root Mean Squared Error for model, Loess and Brass fits

	RMSE		
County Size	Model	Loess	Brass
1,000	0.034	0.187	0.039
5,000	0.027	0.078	0.037
10,000	0.027	0.065	0.042
20,000	0.022	0.052	0.030
100,000	0.013	0.049	0.027

In addition, we also evaluated the relationship between the nominal and actual coverage for the uncertainty intervals produced by the Bayesian model. Coverage is defined as:

$$\frac{1}{G}\sum_{x=1}^{G} 1\left[m_x^* \geq l_x\right] 1\left[m_x^* < r_x\right] \tag{2.22}$$

where G is the number of age groups, m_x^* is the true mortality rate for the xth age group and l_x and r_x the lower and upper bounds of the credible intervals for the xth age group. Coverage at the 80, 90 and 95% levels was considered. Table 2.2 shows the average coverage for the proposed model, fit to 60 counties over 31 years. In general, actual coverage levels are close to the nominal level, indicating that the model is well calibrated. The coverage level tends to decrease as county size increases.

Table 2.2: Nominal versus actual coverage of model uncertainty intervals

	Coverage level (%)		
County Size	80	90	95
1,000	0.871	0.952	0.982
5,000	0.799	0.896	0.940
10,000	0.749	0.853	0.890
20,000	0.744	0.833	0.891
100,000	0.763	0.865	0.901

2.3.2 Application to French *Départements*

We also tested the model on real mortality data, applied to death and population counts by sex in French *départements* from 1975–2008 (INSEE (2015)).[4] The annual life tables were constructed by the *Division des statistiques régionales, locales et urbaines* [Regional, Local and Urban Statistics Division] of the French *Institut national de la statistique et des études économiques* or INSEE [National Institute for Statistics and Economic Studies]. The life tables were built from the vital statistics and census data also collected and processed by INSEE. There are 96 mainland French *départements* ranging in population size from around 35,000 to 1.5 million (for one sex).

We used national France mortality curves from 1975–2008 to form the set of principal components used in estimation. The model was applied to both sexes simultaneously. For illustration, Figs. 2.6 and 2.7 show the observed and estimated log-mortality rates for males in the departments Lozère and Somme in 1975 and 2008. Lozère has a male population of around 38,000, while Somme has a male population of around 275,000. For both *départements*, there is a decrease in mortality rates from 1975 to 2008. As a consequence there are more zero death counts observed in 2008 compared with 1975, corresponding to more uncertainty around estimates in the more recent year. Additionally, there is less uncertainty around the Somme estimates, because the population size is around seven times the population in Lozère.

Figure 2.6: Observed and estimated mortality rates, Lozère, Males, 1975 and 2008.

Once age-specific mortality rates have been estimated, other mortality measures and associated uncertainty can be calculated. Fig. 2.8 below shows life expectancy at birth estimates for males in 2008 by *départements*. Life expectancy is estimated to be highest in areas around Paris and for the Midi-Pyrenees area, and lowest in

[4]We chose to use French *départements* data because, at the time of writing, data for all US counties were not readily available.

Figure 2.7: Observed and estimated mortality rates, Somme, Males, 1975 and 2008.

the northern part of the country, a well-documented pattern (Barbieri and Depledge (2013)).

Uncertainty in life expectancy estimates is also easily obtained. Life expectancy is calculated for each of the posterior samples of age-specific mortality rates. A 95% uncertainty interval is then obtained by calculating the 2.5^{th} and 97.5^{th} percentiles. For example, the estimate for life expectancy at birth for males in Paris in 2008 is 80.5 years (95% UI: [79.4, 81.7]). For a smaller *départements* such as Lozère, the estimate is 79.1 years [77.2, 80.8]. Complete results for all French *départements* are available in an online supplement to this paper.[5] Life expectancy estimates produced by our method are reasonably close to estimates produced by INED (Barbieri and Depledge (2013)), with the average difference in life expectancy at birth across areas and years being around 0.5 years.

Another aspect of the results that could be of interest are the estimated variance parameters. It is assumed that the $\beta_{t,p}$ parameters are normally distributed with mean $\mu_{p,t}$ and variance $\sigma^2_{\beta_{p,t}}$. The variance terms may tell us something about how the spread of mortality outcomes is changing over time. Fig. 2.9 shows the median and 95% credible intervals of the standard deviation parameters associated with β_1 and β_2 over the period of estimation. While there is no discernible trend in $\sigma_{\beta_{1,t}}$, the parameter for the β_2 term, $\sigma_{\beta_{2,t}}$, appears to be increasing over time. The β_2 related to the principal component which alters the relationship of the magnitude of infant and child mortality to mortality at older ages (see Fig. 2.1). An increase in the variance parameter suggests that *départements* are becoming more different over time with respect to child-versus-older mortality. Fig. 2.10 illustrates two *départements* which have relatively high and low values of β_2. For Tarn, β_2 is high, which results from infant mortality being relatively low compared to adult mortality. For Seine-Saint-Denis, the opposite is true.

[5] See `shiny.demog.berkeley.edu/monicah/French/`.

Figure 2.8: Life expectancy estimates for males, 2008 (e_0, years)

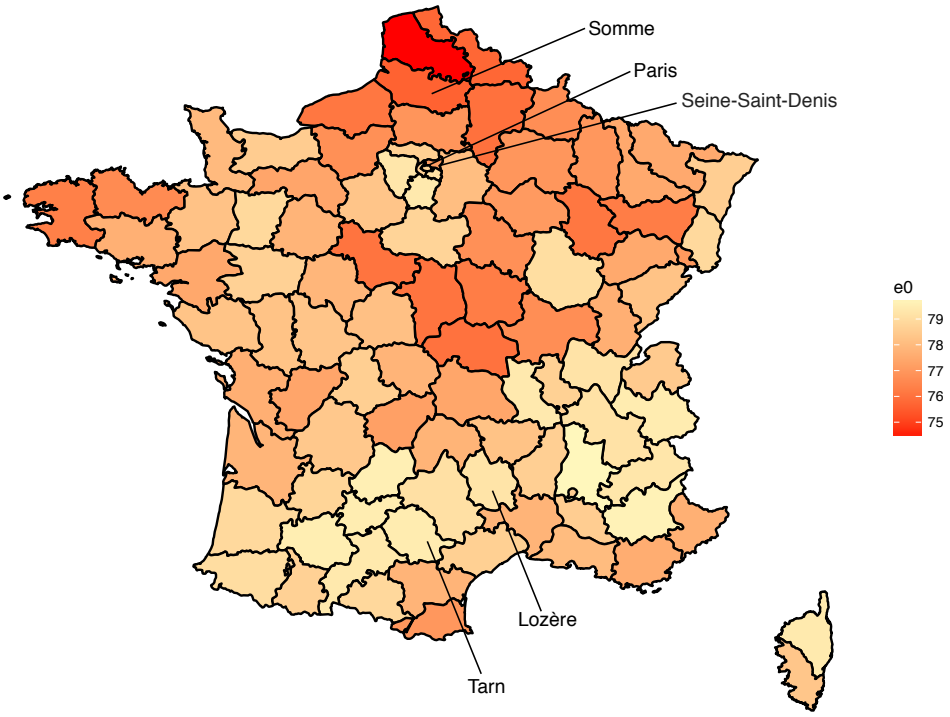
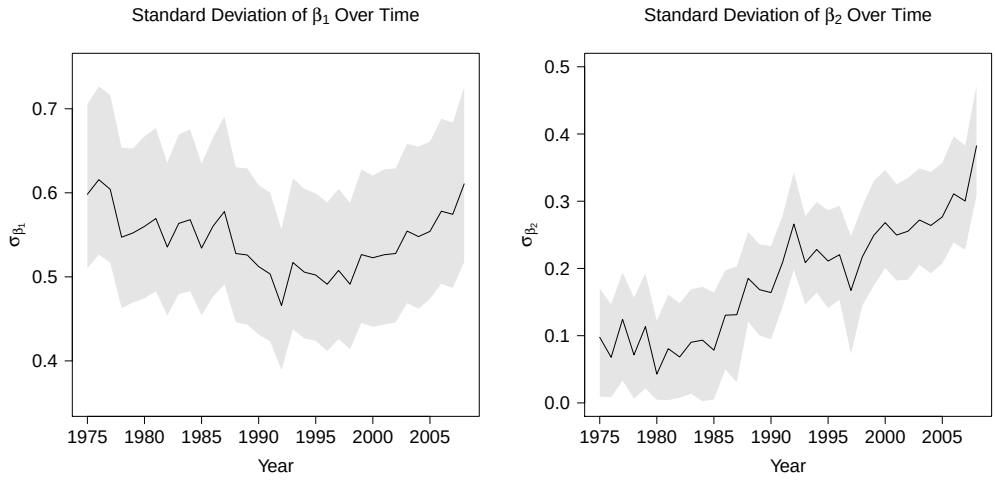


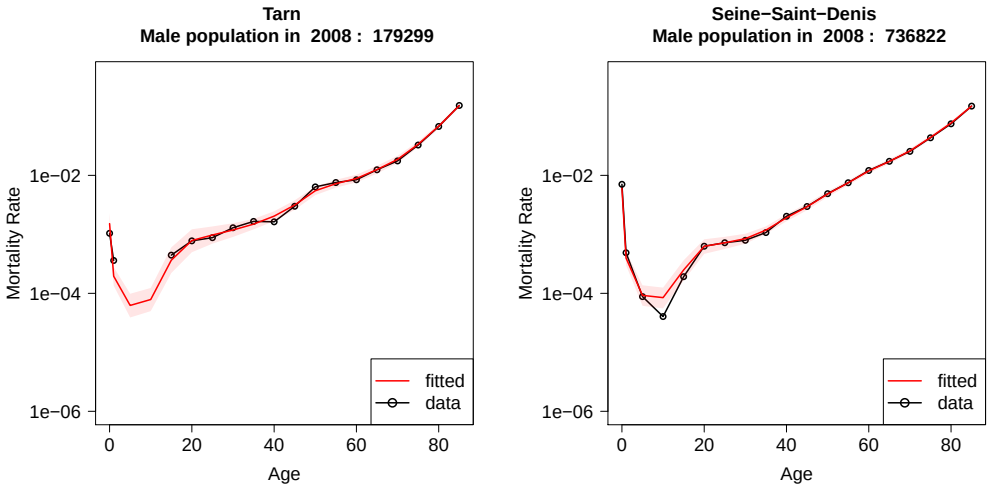
Figure 2.9: Standard deviation of β_1 and β_2 over time



2.4 Conclusion

We presented a novel method to estimate mortality rates by age and sex at the sub-national level. In our approach we build on characteristic age patterns in mortality curves, pooling information across geographic space and smoothing over time within the framework of a Bayesian hierarchical model. When tested against simulated

Figure 2.10: Mortality rates in areas with low (left graph) and high (right graph) values of β_2



data, the model outperformed estimates from a simple Loess smoother and Brass model, especially for areas with smaller population sizes. The uncertainty in our estimates, reflected in the confidence intervals, is well calibrated. An application to real data for France illustrates how various parameter estimates from the model help to assess trends in overall mortality levels and inequalities within a country. The estimates produced by the model have direct applications to the study of subregional health patterns and disparities and how these evolve over time.

The model outlined in the paper is proposed as a general framework for estimating mortality rates in subpopulations. The framework can easily be altered by the user to best suit the situation in which rates are being estimated. For example, the mean parameter values $\mu_{\beta_p,t}$ need not be defined on a state/county basis, but may be defined by a smaller geographic area or based on some other characteristics, such as age distribution or rurality of an area. Additionally, it is possible to alter the random effects to be spatially structured, assuming some correlation in random effects by distance or of adjacent areas.

The focus of this project has been on estimation of past and present mortality trends, rather than future ones. However, forecasting of age- and sex-specific mortality rates in a particular area can be obtained directly from model outputs. The mean parameters $\mu_{\beta_p,t}$ can be projected forward given the assumed linear time trend, which forms a basis to infer other parameters for areas of interest. Given that the relevant variance parameters are also estimated in the process, uncertainty around forecasts can also be inferred.

One of the contributions of our work is methodological. Estimates of mortality measures for small areas, most notably life expectancy, have been proposed previously. For example, Ezzati et al. (2008) used information about number of deaths, together with covariates related to socio-economic status, in order to estimate mortality rates.

Congdon (2014) developed a random effects model to estimate life expectancy for subnational areas. Our method builds on some elements presented in the literature, while incorporating demographic knowledge about regularities in the Lexis surface of mortality rates by age and over time. More specifically, we complemented a state-of-the-art Bayesian hierarchical modeling framework to pool information across space and time, with a classic demographic approach to borrow information across age groups. In particular, our use of principal components of schedules of log-mortality rates is informed by a long tradition of demographic modeling of mortality and can be considered an extension of the Lee-Carter approach (Lee and Carter (1992)).

In this article, we developed a general approach to complement classic demographic modeling ideas within a solid statistical framework. The model that we proposed and tested is quite minimalistic and relies on fairly simple rules for pooling across space and smoothing over time. However, the geographic scale at which spatial pooling will be implemented may depend on specific circumstances of different countries. Likewise, the type of time smoothing may vary. A number of rules can be accommodated within the framework that we propose.

Chapter 3

Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model

3.1 Introduction[1]

In order to evaluate a country's progress in reducing child mortality, it is important to obtain accurate estimates, be able to project mortality levels, and have some indication of the uncertainty in the estimates and projections. In practice, obtaining reliable mortality estimates is often most difficult in developing countries where mortality is relatively high, well-functioning vital registration systems are lacking and the data that are available are often subject to large sampling errors and/or of poor quality. This situation calls for the use of statistical models to help estimate underlying mortality trends.

In recent years, much of the focus on monitoring child mortality has been on assessing changes in the under-five mortality rate (U5MR), which refers to the number of deaths before the age of five per 1,000 live births. The focus was driven by Millennium Development Goal (MDG) 4, which called for a two-thirds reduction in under-five mortality between 1990 and 2015. A report on MDG progress released in 2015 by the United Nations showed that, although this target was not met in most regions of the world, notable progress has been made (UN (2015)). The global U5MR is less than half of its level in 1990, and despite population growth in developing regions, the number of deaths of children under five has declined. Reducing the U5MR continues to be a priority as part of the Sustainable Development Goals (SDG), which replaced the MDGs in 2015. Goal 3 of the SDG includes reducing the

[1]This chapter has been previously published as Alexander and Alkema (2018).

U5MR to at least as low as 25 deaths per 1,000 live births in all countries by 2030 (UN (2017)).

As the U5MR decreases, the share of neonatal deaths, i.e. deaths occurring in the first month, tend to increase. Globally, the estimated share of under-five deaths that were neonatal in 2015 was 45%, a 13% increase from 1990 (IGME (2015)). Indeed, in most regions of the world, the majority of under-five deaths are neonatal; for example, the share is 56% in Developed regions; 51% in Latin America and the Caribbean; and 54% in Western Asia. The share is still less than 50%, however, where the U5MR is relatively high: in Sub-Saharan Africa, the share is only 34%.

The neonatal equivalent to the U5MR is the neonatal mortality rate (NMR), which is defined as the number of neonatal deaths per 1,000 live births. The increasing importance of neonatal deaths in child mortality has warranted increased efforts in monitoring NMR in addition to the U5MR (e.g. Bhutta et al. (2010); Lawn et al. (2004); Lozano et al. (2011)). Goal 3 of the SDGs explicitly includes a neonatal target, with the aim to reduce the NMR to at least as low as 12 deaths per 1000 live births in all countries by 2030 (UN (2017)).

The United Nations Inter-agency Group for Child Mortality Estimation (IGME) publishes estimates of NMR for all 195 UN member countries (IGME (2015)), and these estimates are used to monitor global levels and trends in NMR over time. Up until 2014, IGME used a statistical model to obtain estimates for countries without high-quality vital registration data, which uses U5MR as a predictor Oestergaard et al. (2011). While the method has worked well to capture the main trends in the NMR, it has some disadvantages. Most notably, trends in NMR within a country are driven by the U5MR trends, rather than being specifically influenced by the NMR data.

In this paper, we present a new model for estimating the NMR for countries worldwide, which overcomes some of the concerns with the previous IGME NMR model. We use a penalized splines regression model within a Bayesian hierarchical framework to estimate and project the NMR, and obtain uncertainty around these estimates and projections. In the model, the relationship between NMR and U5MR is used to inform estimates, and the spline regression model is used to capture country-specific trends. From the point of view of modeling mortality levels across countries, a Bayesian approach offers an intuitive way to share information across different countries and time periods, and a data model can incorporate different sources of error into the estimates.

Increases in computational speed as well as the development of suitable numerical methods has enabled a more widespread use of a Bayesian approach in many fields, including population estimation and forecasting (e.g. Alkema and New (2014); Bijak and Bryant (2016); Girosi and King (2008); Raftery et al. (2012); Schmertmann et al. (2014)). The method presented in this paper has similarities to approaches used to estimate other global health indicators, including U5MR (Alkema and New (2014);

You et al. (2015)), maternal mortality Alkema et al. (2016), cause-specific mortality Foreman et al. (2012) and contraceptive prevalence Alkema et al. (2013). In this application, the proposed Bayesian model is flexible enough to be used to estimate the NMR in any country, regardless of the amount and sources of data available. Results were produced for 195 countries for at least the years 1990–2015, which covers the MDG period of interest, using a dataset with almost 5000 observations from various data sources.

The remainder of the paper is structured as follows. Firstly, the dataset and model are summarized in the next two sections. Some key results are then highlighted, including model validation results, followed by a discussion of the work and possible future avenues. Additional details about the model are provided in the Appendix.

3.2 Data

There is large variability in the availability of data on neonatal mortality. Broadly there are three main data sources for the neonatal mortality rate: vital registration (VR) systems; sample vital registration (SVR) systems; and survey data. Data for a particular country may come from one or several of these sources, and the source type may vary over time. Table 3.1 summarizes the availability of data by source type.

3.2.1 Source types

Data from VR systems are derived directly from the registered births and deaths in a country. The observed NMR for a particular country and year is the number of registered deaths within the first month divided by the number of live births. Because VR data are based on the records from the whole population, they are usually high quality compared to other sources. Most developed countries have VR data available. SVR systems refer to vital registration statistics that are collected on a representative sample of the broader population.

NMR observations can also be derived from data collected in surveys, if women are asked to list a full history of all births (and possible deaths) of her children. A retrospective series of NMR observations can then be derived using the birth histories. A total of 72% of the survey data series contained in the database (Table 1) have microdata available and it is possible to estimate the sampling error associated with each of the observations. For the remaining 28%, data come from summary reports and preliminary releases; as such there is not enough information to calculate the sampling errors from the data. For these series, values for sampling errors are imputed (see Section 3.3.4). All mortality rates, ratios of mortality rates, and corresponding standard errors were calculated from the survey microdata using the

software 'CMRJack' Pedersen and Liu (2012), a software package that produces mortality estimates and standard errors for surveys with complete birth histories or summary birth histories. Estimates are obtained based on the methodology outlined in Pedersen and Liu (2012). The retrospective time period covered by mortality estimates is optimized to capture short-term fluctuations while still ensuring that the estimates have a coefficient of variation of less than 10%.

The majority of survey data come from Demographic and Health Surveys (DHS) (Table 3.1). The category 'Other DHS' refers to non-standard DHS, that is, Special Interim and National DHS, Malaria Indicator Surveys, AIDS Indicator Surveys and World Fertility Surveys (WFS). National DHS are surveys in DHS format that are run by a national agency, rather than the external DHS agency. The Multiple Indicator Cluster Survey (MICS), developed by UNICEF in 1990, was originally designed to address trends in goals from the World Summit for Children, and has since focused on assessing progress towards the relevant MDG indicators. The 'Other' category includes surveys such as the Pan-Arab Project for Family Health and the Reproductive Health Surveys.

3.2.2 Data availability

Data availability varies by country and by year. For most developed countries, a full time series of VR data exists. For other countries with VR data, the time series is often incomplete and is supported by other sources of data. Of the 105 countries where VR data are available, 44 countries have incomplete VR times series. For some smaller countries with VR data, observations were combined to avoid issues with erratic trends due to large stochastic variance (see Appendix A.2.3 for more details). SVR data are only available for Bangladesh, China and South Africa. Most developing countries have no vital registration systems and so observations of the NMR are derived entirely from surveys. A total of 12 countries had no available data.

In terms of data inclusion, we follow the same inclusion exclusion rules as the UN IGME-estimated U5MR (IGME (2015)). These exclusion rules are based on external information which suggests some NMR observations are unreliable, due to, for example, poor survey quality or under coverage of VR systems. A total of 16% of the 4,678 observations were excluded.

Fig. 4.2.1 illustrates examples of the data available for four countries. The shaded area around the observations has a width of two times the sampling error (for survey data) or stochastic error (for VR data). The NMR for Australia (Fig. 3.1a), as calculated from the full VR data time series, has a trend over time which is relatively regular and the uncertainty is low. Data for Sri Lanka indicate NMR are roughly five times as high as Australia. Data are available from 1950, but the VR data series is incomplete. The rest of the data come from the WFS, DHS and National

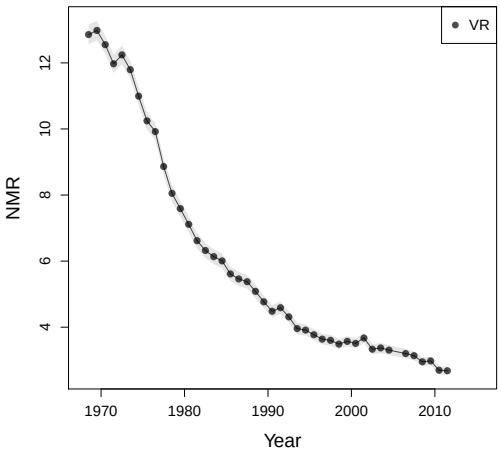
Table 3.1: Summary of the NMR data availability by source type. The totals include observations that were excluded from the estimation.

Source	Sampling Errors	No. of Series	No. of Countries	No. of Obs	No. of Country-years
VR	Calculated	105	105	2607	2607
SVR	Calculated	3	3	79	78
DHS	Reported	239	81	1212	934
DHS	Unreported	16	15	50	48
Other DHS	Reported	52	42	251	251
Other DHS	Unreported	26	21	78	75
MICS	Reported	16	14	81	73
MICS	Unreported	12	12	49	46
Others	Reported	24	16	119	111
Others	Unreported	72	36	152	151

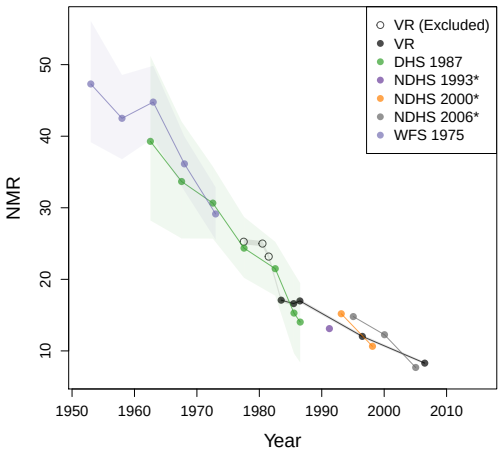
DHS. There are multiple estimates for some years, and the uncertainty around the estimates varies by source and year. The uncertainty around the VR data is much less than for the survey data. The National DHS series do not have estimates of sampling error. Iraq (Fig. 3.1c) has no VR data, and the estimates are constructed from MICS and two other surveys: the Infant and Child Mortality and Nutrition Survey, and the Child and Maternal Mortality Survey. Again, there are multiple estimates for some time points, and uncertainty level and availability varies. Finally, Vanuatu (Fig. 3.1d) has only three observation points from one National DHS.

3.3 Method

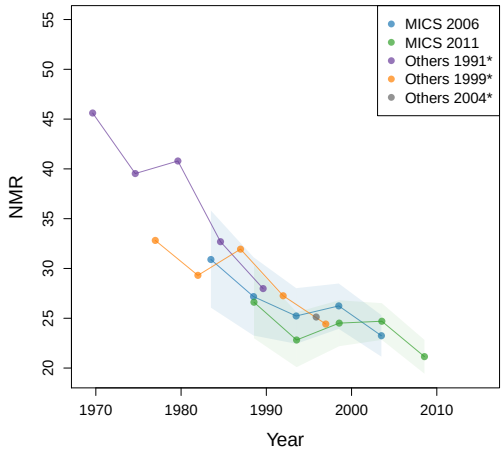
The aim is to produce estimates of the NMR for all countries in the world, and report the associated uncertainty around the estimates. The model needs to be flexible enough to estimate NMR in a variety of situations, as illustrated in Fig. 4.2.1. The estimates should follow the data closely for countries with reliable data and low uncertainty. On the other hand, the model estimates need to be adequately smooth in countries with relatively large uncertainty and erratic trajectories. The model also needs to be able to estimate NMR over the period 1990-2015 for *all* countries, including those countries where there are limited or no data available. To achieve these goals, our proposed model utilizes the relationship between the U5MR and NMR: as the level of U5MR decreases, the proportion of deaths under five that are neonatal tends to increase. In addition, the model also allows for country-specific effects and time trends to capture data-driven trends in data-rich countries. The term 'data-driven' refers to a model set-up where the NMR estimates over time are explicitly influenced by temporal changes in the NMR data. This is in contrast to a model where temporal changes in NMR estimates are driven by trends in U5MR only, as was the case with the previous IGME model Oestergaard et al. (2011).



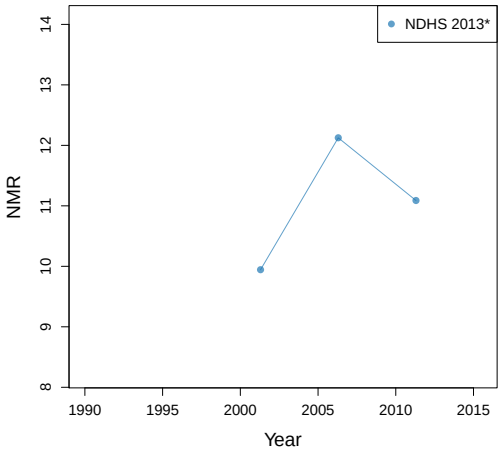
(a) Australia



(b) Sri Lanka



(c) Iraq



(d) Vanuatu

Figure 3.1: Neonatal Mortality Data (deaths per 1,000 births) for four selected countries. The different colored circles represents different data sources, as described in the plot legends. An open circle indicates that the observation was excluded from the analysis. The shaded area around the data series represents the stochastic error (in the case of VR data) and sampling error (in the case of survey data). Survey data series that did not have reported sampling errors do not have a shaded region on the plots, and are marked with an asterisk (*) in the legends.

In the NMR model, we use country-year specific U5MRs as explanatory variables and also to obtain final estimates of NMR. All estimates of U5MR used in the model were obtained from the UN IGME (IGME (2015)).

3.3.1 Model overview

Write $N_{c,t}$ and $U_{c,t}$ as the NMR and U5MR for country c at time t, respectively, with $U_{c,t}$ given by the IGME U5MR estimate for that country-year. Note that $N_{c,t}$ and $U_{c,t}$ are always expressed in units of deaths per 1,000 live births. We explain the model set-up in terms of the ratio

$$R_{c,t} = \frac{N_{c,t}}{U_{c,t} - N_{c,t}},$$

which refers to the true ratio of neonatal deaths compared to deaths in months 2 to 60. We constrain $R_{c,t} > 0$ such that $0 \leq \frac{N_{c,t}}{U_{c,t}} \leq 1$ to guarantee that NMR estimates are not greater than U5MR estimates. The true ratio $R_{c,t}$ is modeled as follows:

$$R_{c,t} = f(U_{c,t}) \cdot P_{c,t}, \tag{3.1}$$

where $f(U_{c,t})$ is the overall expected ratio given the current level of U5MR and $P_{c,t}$ is a country-specific multiplier to capture deviations from the expected relationship.

The observed ratio $r_{c,i}$, which refers to the i-th observation of the ratio in country c, is expressed as a combination of the true ratio and some error, i.e.

$$\begin{aligned} r_{c,i} &= R_{c,t[c,i]} \cdot \epsilon_{c,i} \\ \implies \log(r_{c,i}) &= \log(R_{c,t[c,i]}) + \delta_{c,i} \\ &= \log f(U_{c,t[c,i]}) + \log P_{c,t[c,i]} + \delta_{c,i} \end{aligned} \tag{3.2}$$

for $c = 1, 2, \ldots, C$ and $i = 1, \ldots, n_c$, where $C = 195$ (the total number of countries) and n_c is the number of observations for country c. The index $t[c, i]$ refers to the observation year for the i-th observation in country c, $\epsilon_{c,i}$ is the error of observation i and $\delta_{c,i} = \log(\varepsilon_{c,i})$. Note that throughout the paper, log refers to the natural logarithm.

The following sections explain how we chose to model the expected ratio $f(U_{c,t})$, the country-specific multiplier $P_{c,t}$, and the error term $\delta_{c,i}$. Other aspects of the method, including the projection method, estimation for countries with no data and crisis and HIV/AIDS adjustments are detailed in the Appendix.

3.3.2 Global relationship with U5MR

The first step in modeling the ratio of neonatal to non-neonatal deaths is to find an appropriate function $f(U_{c,t})$ in Eq. 3.1, which captures the expected value of the ratio

given the current level of U5MR. We modeled $f(U_{c,t})$ on the log scale. Fig. 3.2 shows a scatter plot of log-transformed observed ratios $\log(r_{c,i})$ versus $\log(U_{c,t[c,i]})$. The relationship between the two variables appears to be relatively constant up to around $\log(U_{c,t}) = 3.5$, after which point the log ratio decreases linearly with decreasing $\log(U_{c,t})$. Given this observed relationship, $\log f(U_{c,t})$ is modeled as follows:

$$\log f(U_{c,t}) \quad = \quad \begin{cases} \beta_0 & \text{for } U_{c,t} \leq \theta, \\ \beta_0 + \beta_1 \cdot (\log(U_{c,t}) - \log(\theta)) & \text{for } U_{c,t} > \theta. \end{cases}$$

This implies, below a cutpoint θ, $\log f(U_{c,t})$ is modeled as a constant β_0. Above the cutpoint, $\log f(U_{c,t})$ is represented linearly as a function of $\log(U_{c,t})$ with slope β_1.

The fitted relationship between the ratio and the level of U5MR from the NMR model, $\log f(U_{c,t})$, is illustrated in Fig. 3.2. The posterior median estimate for the cutpoint θ is 34.2 deaths per 1,000 births (90% CI: [33.7, 34.5]). At U5MR levels that are higher than θ, the β_1 coefficient suggests that a 1% increase in the U5MR leads to a 0.65% decrease (90% CI: [0.61, 0.71]) in the ratio $R_{c,t}$. The fitted line is quite similar in shape to the loess curve fitted to the data, shown by the red line in Fig. 3.2.

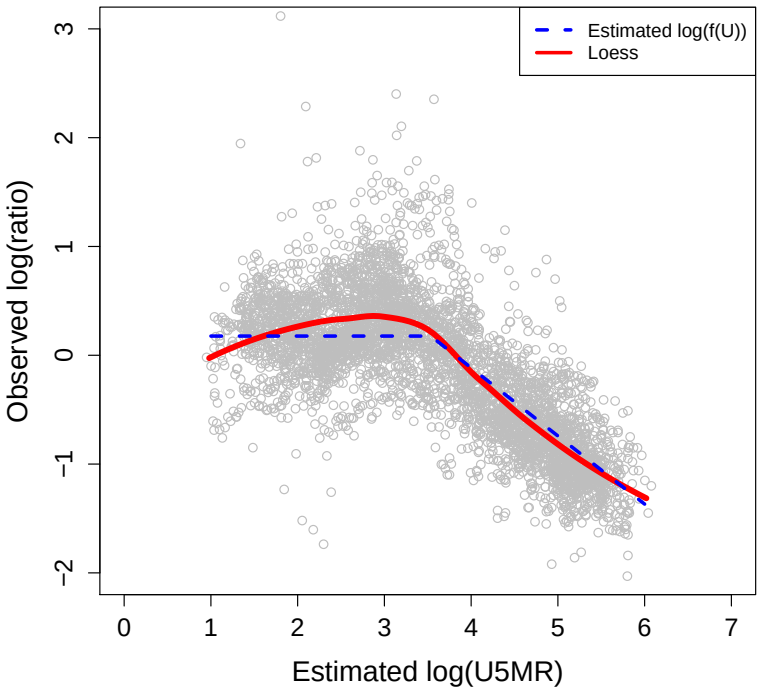


Figure 3.2: Observed and estimated relation between the ratio of neonatal and non-neonatal deaths and under-five mortality. Observations $\log r_{c,i}$ are displayed with grey dots and plotted against $\log U_{c,t[c,i]}$. The Loess fit to the observations is shown in red, and the estimated relation (function $f(U_{c,t})$) is added in blue (dashed line).

3.3.3 Country-specific multiplier

This section details how the country-specific multiplier $P_{c,t}$ is modeled. Although there is a relationship between the neonatal ratio and U5MR at the aggregate level, the relationship between $R_{c,t}$ and $U_{c,t}$ is likely to differ by country. For instance, some countries may have higher or lower levels of NMR than what is expected given the level of U5MR. In addition, within a particular country, the relationship between NMR and U5MR may not be constant over time, so the model should be flexible enough to also allow for temporal changes. This is the purpose of the country-specific term $P_{c,t}$ in Eq. 3.1: to capture data-driven differences across countries and also within countries over time.

The country-year multiplier $P_{c,t}$ was modeled on the log scale with a basis-splines (B-splines) regression model:

$$\log(P_{c,t}) = \sum_{k=1}^{K_c} B_{c,k}(t)\alpha_{c,k},$$

where $B_{c,k}(t)$ refers to the kth B-spline function for country c evaluated at time t and $\alpha_{c,k}$ is the k-th splines coefficient for country c. The B-splines $B_{c,k}(t)$, which are illustrated in Fig. 3.3a for Nigeria, were constructed using cubic splines. In the figure, each $B_{c,k}(t)$ is represented in a different color at the bottom. Spline placement is determined by knot points, indicated by gray dotted vertical lines. Knot points occur where the spline function is at its maximum. Country c has a total of K_c knot points defined by $t_1 < t_2 < \cdots < t_{K_c}$. K_c is the number of B-splines needed to cover the period up to 2015 and back to 1990 or the start of the observation period, whichever is earlier. In terms of knot spacing, the same interval length of 2.5 years was used in each country regardless of the number or spacing of observations. The consistent interval length was chosen to be able to exchange information across countries about the variability in changes between spline coefficients. Knot placement was determined by placing one knot half an interval before the most recent observation year in each country. Because the most recent observation year differs by country, the splines $B_{c,k}(t)$ also differ by country.

The fitting of the country-specific multiplier $\log P_{c,t}$ for Nigeria is also illustrated in Fig. 3.3a. The splines regression for $\log P_{c,t}$ captures any pattern in the data on the log scale once the global relation between ratio and U5MR, as expressed by $f(U_{c,t})$, have been taken into account. As such, the y-axis in Fig. 3.3a refers to, on the log scale, the difference between the observed data points and their expected level given the U5MR, i.e. $\log(r_{c,i}) - \log f(U_{c,t[c,i]})$. The different colored dots connected with lines represent residual data points from different sources available for Nigeria. The estimated $\log P_{c,t}$ at a particular time point t is given by a linear combination of the $B_{c,k}$ at point t and the estimated coefficients $\alpha_{c,k}$.

The splines regression model for $\log(P_{c,t})$ is very flexible in order to be able to capture patterns in the data. However, in situations where the data are sparse, limited

information on a subset of the spline coefficients $\alpha_{c,k}$ can result in an implausible fit. We impose smoothness on the fits by penalizing differences in adjacent spline coefficients $\alpha_{c,k}$. This is referred to Penalized splines, or P-splines regression (Eilers and Marx (1996), Currie and Durban (2002)).

In the P-splines regression, spline coefficients are modeled as a combination of an overall mean value λ_c and $K_c - 1$ first-order differences

$$\boldsymbol{\varepsilon}_c = (\alpha_{c,2} - \alpha_{c,1}, \alpha_{c,3} - \alpha_{c,2}, \ldots, \alpha_{c,K_c} - \alpha_{c,K_c-1}).$$

The λ_c can be interpreted as a country-specific intercept, representing deviations in the level from the overall global relationship between $R_{c,t}$ and $U_{c,t}$. We model the λ_c's centered at zero:

$$\lambda_c \sim N(0, \sigma_\lambda^2).$$

The $\boldsymbol{\varepsilon}_c$ term represents fluctuations around the country-specific intercept. These fluctuation terms allow for the $P_{c,t}$ term to be influenced by the changes in the level of the underlying data. The fluctuations are modeled as:

$$\varepsilon_{c,k} \sim N(0, \sigma_{\varepsilon_c}^2). \tag{3.3}$$

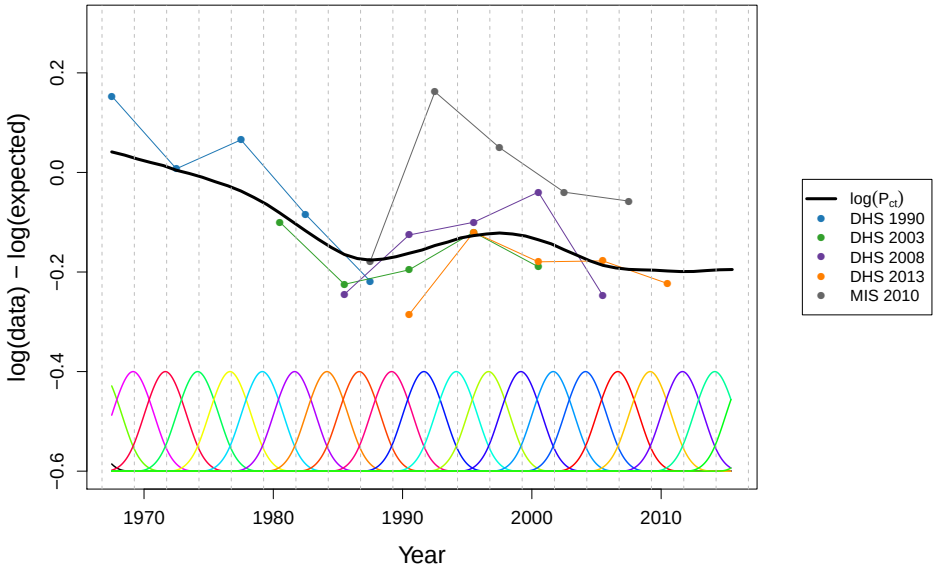
The variance $\sigma_{\varepsilon_c}^2$ essentially acts as a country-specific smoothing parameter. The smoothness of a particular country's trajectory depends on the regularity of the trend in the data and also the measurement errors associated with the data points. As $\sigma_{\varepsilon_c}^2$ decreases, the fluctuations go to zero, and the $\alpha_{c,k}$'s become a country-specific intercept with no change over time. The $\sigma_{\varepsilon_c}^2$'s are modeled hierarchically:

$$\log(\sigma_{\varepsilon_c}^2) \sim N(\chi, \psi_\sigma^2), \tag{3.4}$$

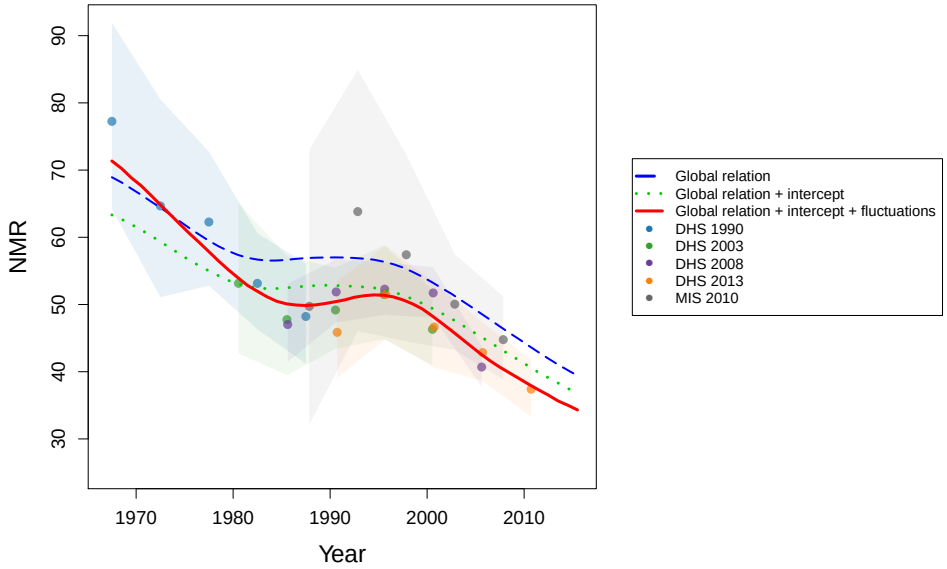
where e^χ can be interpreted as a 'global smoothing parameter' and ψ_σ^2 reflects the across-country variability in smoothing parameters. The hierarchical structure of the model allows information on the amount of smoothing to be shared across countries. The countries with fewer data points and thus less information about the level of smoothness borrow strength from countries with more observations.

The effect of including the country-specific term is shown in Fig. 3.3b for Nigeria. The available data series are shown by the points, and the shaded area around those points represents their associated sampling error. The blue solid line illustrates the fit from the global relation, i.e. $f(U_{c,t})$. The green line illustrates from the global relation and country-specific intercept, i.e. a combination of $f(U_{c,t})$ and λ_c. Note that this line has the same shape as the blue line, but has been lowered. The red solid line shows the final fit after inclusion of the fluctuations i.e. $f(U_{c,t})$, λ_c and $\boldsymbol{\varepsilon}_c$. This allows the fitted trajectory to be more influenced by the data.

The choice to model the country-specific terms $P_{c,t}$ using P-splines regression was motivated by previous work on mortality modeling and forecasting (e.g. Alkema and New (2014), Currie et al. (2004); D'Amato et al. (2011)). In practice, there are many different ways to model and smooth the country-specific term $P_{c,t}$. For example, $P_{c,t}$ could have been modeled using an autoregressive or autoregressive-moving average



(a) Estimate of $\log P_{c,t}$ for Nigeria using splines regression. The y-axis is (on log scale) the difference between the observed data points and the expected level (given by $f(U_{c,t})$). The different colored dots/ lines represent data from different sources available. Each basis spline is represented in a different color at the bottom of the figure. These have been scaled vertically for display purposes. The gray dotted vertical lines indicate knot positions (every 2.5 years).



(b) The three components of fit: The blue dashed line illustrates the fit from the global relation, i.e. $f(U_{c,t})$. The green dashed line illustrates from the global relation and country-specific intercept, i.e. a combination of $f(U_{c,t})$ and λ_c. The red solid line shows the final fit after inclusion of the fluctuations i.e. $f(U_{c,t})$, λ_c and $\boldsymbol{\varepsilon}_c$.

Figure 3.3: Illustration of splines regression and three fit components for Nigera

(ARMA) process, similar to the modeling approaches for global maternal mortality (Alkema et al. (2016)) and contraceptive prevalence (Alkema et al. (2013)). However, the use of a splines basis, which results in a regression function that is twice differentiable, gives estimates of $P_{c,t}$ that are relatively smooth compared to an ARMA-based approach. In addition, the P-splines regression approach was chosen for consistency with the current model used by IGME to estimate U5MR.

3.3.4 Data model

Eq. 3.2 indicates that the observed ratio $r_{c,i}$ is modeled on the log-scale as the true ratio $R_{c,t}$ plus some error term $\delta_{c,i}$. We model this error term $\delta_{c,i}$ differently based on the source of the data of the i-th observation. The model imposed on $\delta_{c,i}$ is called the 'data model'.

VR data

For VR data series, the error term $\delta_{c,i}$ is modeled as

$$\delta_{c,i} \sim N(0, \tau_{c,i}^2),$$

where $\tau_{c,i}$ is the stochastic standard error. These are obtained based on standard assumptions about the distribution of deaths in the first month of life. Details are given in Appendix B1. Note that SVR data are modeled the same as VR data, but the $\tau_{c,i}$ term refers to the sampling error.

Non-VR data

For the non-VR data, the error term δ_i is modeled as

$$\delta_{c,i} \sim N(0, \nu_{c,i}^2 + \omega_{s[c,i]}^2),$$

where $\nu_{c,i}$ is the sampling error and $\omega_{s[c,i]}$ is non-sampling error of the series type s of observation i in country c. Non-sampling error variances are estimated separately for each of the series types listed in Table 3.1: DHS, Other DHS, MICS, and Others. The distinction by series type was made to allow for the possibility that a particular survey may be run in a similar fashion across different countries, and as such may display similar characteristics in terms of non-sampling error.

Sampling error variances were reported for the majority of the non-VR observations (see Table 3.1). For those observations (c, i) where sampling error was not reported, it was imputed based on the median value of all observed sampling errors of series type $s[c, i]$ within the group-size category of country c. A country was categorized

as 'small' if the annual number of births was in the lowest quartile of all countries (corresponding to a maximum of around 25,000 births per year). The distinction between small and other-sized countries was made due to the differences in observed standard errors. The imputed values for missing standard errors for each size category and series type are shown in Table 3.2.

Table 3.2: Values imputed for missing standard errors for survey data by series and country size category.

	Country size category	
Series type	Other	Small
DHS	0.13	0.26
MICS	0.16	0.21
Other DHS	0.14	0.24
Others	0.16	0.22

3.3.5 Obtaining the final estimates

The model described above produces estimates of $\log(R_{c,t})$. The corresponding estimate of $N_{c,t}$ is obtained by transforming the ratio and combining it with $U_{c,t}$:

$$N_{c,t} = \text{logit}^{-1}(\log(R_{c,t})) \cdot U_{c,t},$$

because

$$\text{logit}\left(\frac{N_{c,t}}{U_{c,t}}\right) = \log\left(\frac{N_{c,t}}{U_{c,t} - N_{c,t}}\right) = \log(R_{c,t}).$$

The ratio estimates are recombined with IGME estimates of $U_{c,t}$. However, using only the median estimates of $U_{c,t}$ does not take into account the level of uncertainty in the $U_{c,t}$ estimates and correspondingly under-represents the level of uncertainty in the $N_{c,t}$. As such, the $N_{c,t}$ estimates were generated by randomly combining posterior draws of $\text{logit}^{-1}(\log(R_{c,t}))$ and of $U_{c,t}$. The result is a series of trajectories of $N_{c,t}$ over time. The best estimate is taken to be the median of these trajectories and the 5th and 95th percentiles are used to construct 90% credible intervals.

3.3.6 Computation

The hierarchical model detailed in the previous sections is summarized in Appendix A.1. The model was fitted in a Bayesian framework using the statistical software R. Samples were taken from the posterior distributions of the parameters via a Markov Chain Monte Carlo (MCMC) algorithm. This was performed through the use of JAGS software Plummer (2003).

In terms of computation, three chains with different starting points were run with a total of 20,000 iterations in each chain. Of these, the first 10,000 iterations in each

chain were discarded as burn-in and every 10th iteration after was retained. Thus 1,000 samples were retained from each chain, meaning there were 3,000 samples retained for each estimated parameter.

Trace plots were checked to ensure adequate mixing and that the chains were past the burn-in phase. Gelman's $\hat{R}$ Gelman and Rubin (1992) and the effective sample size were checked to ensure a large enough and representative sample from the posterior distribution. The value of $\hat{R}$ for all parameters estimated was less than 1.1.

3.4 Results

Estimates of NMR were produced for the 195 UN member countries for at least the period 1990–2015, with periods starting earlier if data were available. In this section some key results are highlighted. Results are also compared to those produced by the method previously used by the IGME.

The estimated global relation (Table 3.3) suggests that the relationship between the ratio and U5MR is constant up to a U5MR of 34.2 (90% CI: [33.7, 34.5]) deaths per 1,000 births, the ratio of neonatal to other child mortality is constant at around 1.20 (90% CI: [1.03, 1.25]). This is equivalent to saying the proportion of deaths under-five that are neonatal is constant at around 54% (90% CI: [50, 55]). Above a U5MR of 34, the estimated coefficient suggests that, at the global level, a 1% increase in the U5MR is associated with a 0.65% (90% CI: [0.61, 0.70]) decrease in the ratio.

Table 3.3: Estimates for parameters in global relation

	Median	90% CI
β_0	0.18	(0.03, 0.22)
β_1	-0.65	(-0.70, -0.61)
θ	34.2	(33.7, 34.5)

3.4.1 Results for selected countries

The fits for the four countries illustrated in Section 2 are shown in Fig. 3.4. In the figures, the blue dashed line represents the expected level of NMR given the country's U5MR. The solid red line and associated shaded area represents model estimates and 90% uncertainty intervals. For Australia (Fig. 3.4a), the estimates follow the data closely, given the small uncertainty levels around the data. There has been a steady decrease in NMR since 1970. In earlier time periods, the level of NMR was higher than the expected level (that is, the solid red line is higher than the blue dashed line). This switched in the 1980s and 1990s, and more recently, the estimated and expected levels are close.

For Sri Lanka (Fig. 3.4b), the estimates of NMR are informed by the combination of VR and survey data. The VR has a greater influence on the trajectory because of the smaller associated standard errors. In the earlier years, the uncertainty intervals around the estimate are larger due to the higher uncertainty of the data. There is a small spike in the estimate in the year 2004, which is a tsunami-related crisis adjustment.

No VR data were available for Iraq (Fig. 3.4c), and the larger sampling errors around the survey data have led to relatively wide uncertainty intervals over the entire period. This is in contrast to Sri Lanka, where uncertainty intervals became more narrow once VR data were available. The larger sampling errors in Iraq have also led to a relatively smooth fit (high value of the smoothing parameter), and the shape of the trajectory essentially follows the shape of the expected line.

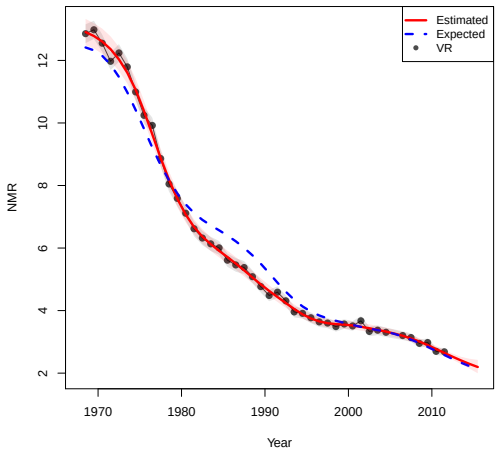
For Vanuatu (Fig. 3.4d), the trajectory is driven by the expected trajectory given Vanuatu's trend in U5MR. The available data determine the country-specific intercept for Vanuatu, which is lower than the expected level. However, the relative absence of data for this country means that the uncertainty around the estimates is high.

3.4.2 Outlying countries

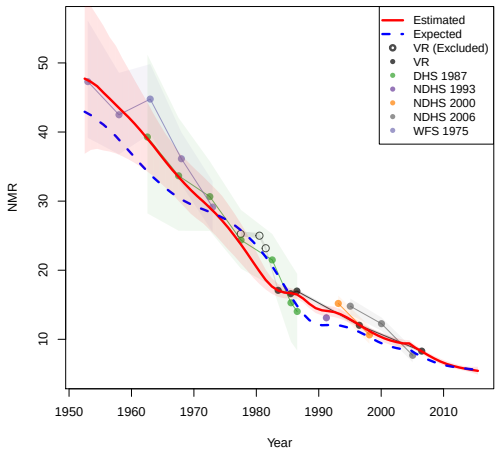
The set-up of the model allows for the comparison of the estimated level of NMR to the expected level of NMR given the U5MR. The expected level as predicted by U5MR is an estimation with $f(U_{c,t})$ only (without the country-specific effect, $P_{c,t}$). We define a country to be outlying if the the estimated NMR in 2015 was significantly higher or lower than the expected level by at least 10%. That is, the ratio of estimated-to-expected was at least 1.1 or less than 0.9 in 2015, and the 95% credible interval does not contain 1. Fig. 3.5 illustrates these countries, and the values of estimated-to-expected in 1990 and 2015.

Countries that have a lower-than-expected NMR include Japan, Singapore and South Korea, and some African countries such as South Africa and Swaziland. Countries that have a higher-than-expected NMR include several Southern Asia countries, such as Bangladesh, Nepal, India and Pakistan. The former Yugoslavian countries Croatia, Bosnia and Herzegovina, and Montenegro also have higher-than-expected NMR.

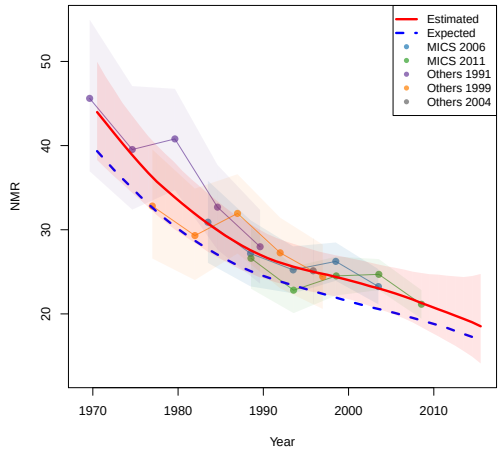
Fig. 3.6 shows estimates through time for two contrasting countries, Japan, which has lower-than-expected NMR and India, with higher-than-expected NMR. In each of the figures, the red line represents the estimated fitted line (with 90% CIs). The blue line represents the expected level, which can be interpreted as the expected level of NMR in a particular year as predicted by the level of U5MR. The gap between the expected and estimated NMR is being sustained through time for Japan, and



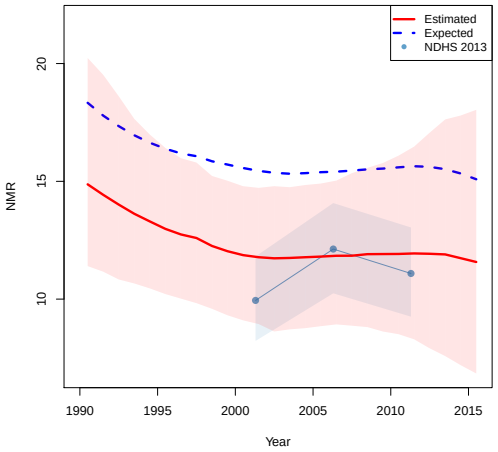
(a) Australia



(b) Sri Lanka



(c) Iraq



(d) Vanuatu

Figure 3.4: Observed and estimated neonatal mortality (deaths per 1,000 births) for selected countries. The blue dashed line represents the expected level of NMR given the country's U5MR. The solid red line and associated shaded area represents model estimates and 90% uncertainty intervals.

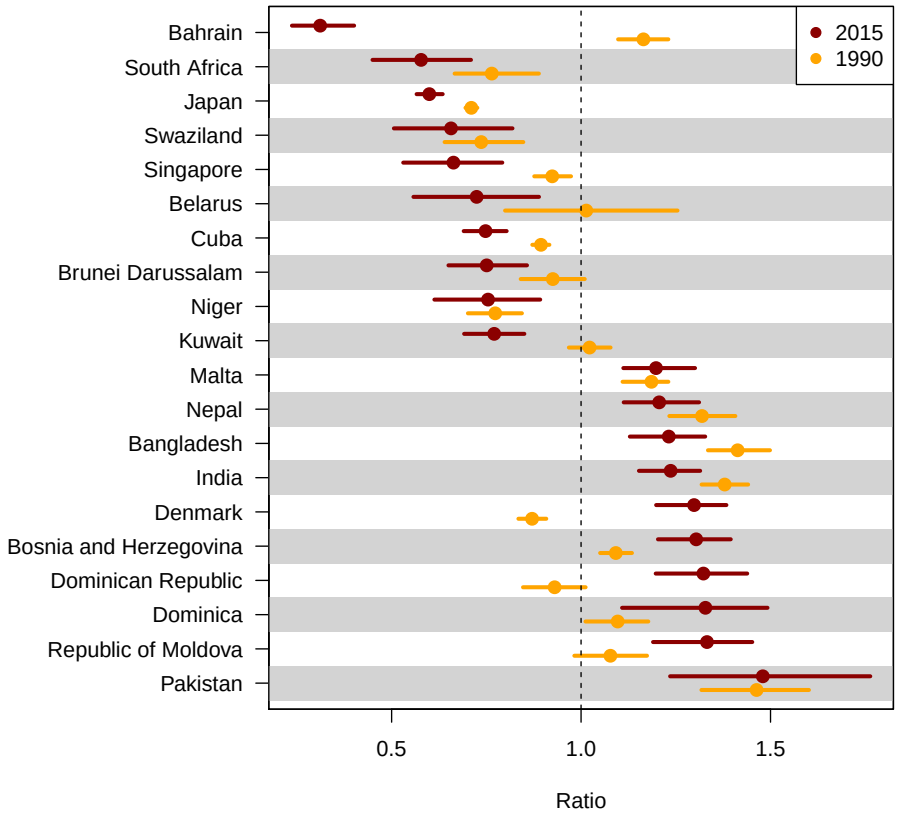


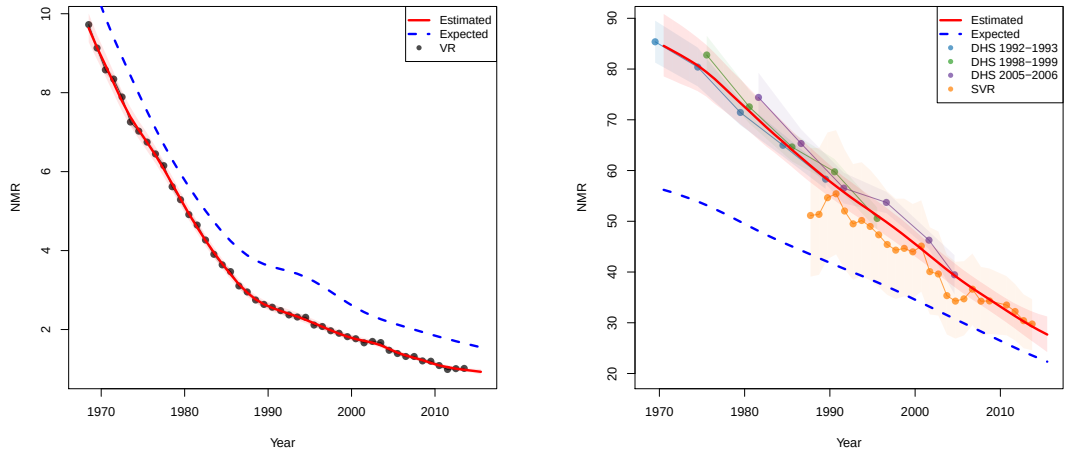
Figure 3.5: Ratio of estimated NMR to the expected NMR given U5MR for outlying countries. A country is outlying if the the estimated NMR in 2015 was significantly higher or lower than the expected level given U5MR by at least 10%. The dot shows the median estimate, and the lines give the 95% uncertainty interval.

has widened since the 1970s. The change in NMR levels for India has been dramatic. Not only is the current NMR around 30% of what it was in 1970, the discrepancy between the expected and estimated levels has decreased through time.

3.4.3 Smoothing

The smoothness of the fluctuations $\sigma^2_{\varepsilon_c}$ is modeled hierarchically, assuming a log-normal distribution with a mean parameter χ (see Eq. 3.4). Smoothing parameters can also be expressed in terms of precision, $1/\sigma^2_{\varepsilon_c}$; Fig. 3.7 shows the distribution of estimated precisions for all countries. The larger the value of the smoothing parameter (precision), the smoother the fit. The estimate of the mean smoothing parameter was around 59 (90% CI: [43, 79]).

Larger values of smoothing parameters were estimated for countries that had no available VR data but many observations from survey data. Senegal, which had



(a) Japan (b) India

Figure 3.6: Two example outlying countries. The level of NMR is lower than the expected as predicted by U5MR in Japan, while the level of NMR in India is higher than expected.

the highest smoothing parameter at a value of 582 (90% CI: [113, 4000]), had a total of 55 observations over a 45-year period (Fig. 3.8a). The effect of having many observations with relatively large standard errors is a relatively smooth fit. In contrast, one of the smallest smoothing parameters occurred for Cuba, at around 4 (90% CI: [2, 7]). Cuba is a country with good quality VR-data, which has relatively small standard errors. This means the fitted line follows the data more closely (Fig. 3.8b).

3.4.4 Comparison with existing IGME model

It is useful to compare the results of this new model to the NMR results from the model previously used by the IGME. The previous model is described in Oestergaard et al. (2011). In this method, NMR estimates for countries with complete VR series are taken directly from the data. For countries without a complete VR series, a multilevel model is fit using U5MR as a predictor, with a quadratic relationship specified. In addition, the model allows for country-level and region-level random effects:

$$\log(NMR_{c,t}) = \underbrace{\alpha_0 + \beta_1 \log(U_{c,t}) + \beta_2 (\log U_{c,t})^2}_{\log(f(U_{c,t}))} + \underbrace{\alpha_{country[i]} + \alpha_{region[i]}}_{\log(P_{c,t})}$$

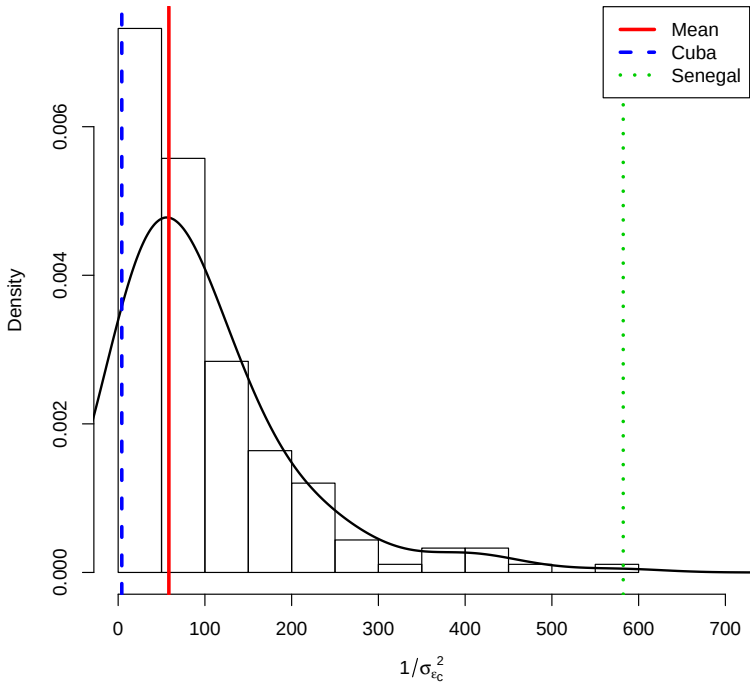
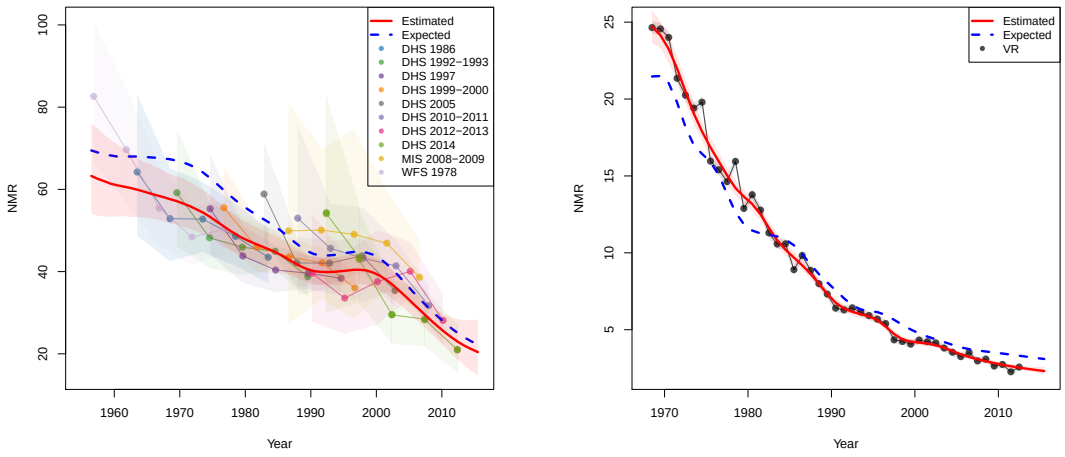


Figure 3.7: Distribution of estimated precisions ($1/\sigma^2_{\varepsilon_c}$) relating to smoothing parameters for all countries. The red solid line represents the mean value of all precisions. The blue dashed line is the estimated smoothing parameter for Cuba (relatively little smoothing), while the dotted line is the estimated smoothing parameter for Senegal (relatively high smoothing

.



(a) Senegal

(b) Cuba

Figure 3.8: Example countries with relatively high smoothing (Senegal) and relatively low smoothing (Cuba).

For comparison, the new model is:

$$\begin{aligned}\log(R_{c,t}) &= \underbrace{\beta_0 + \beta_1 \cdot (\log(U_{c,t}) - \log(\theta))_{[\log(U_{c,t})>\log(\theta)]}}_{\log(f(U_{c,t}))} \\ &+ \underbrace{\sum_{k}^{K_c} B_{c,k}(t)\alpha_{c,k}}_{\log(P_{c,t})}\end{aligned}$$

The existing model is similar in that it estimates NMR as a function of U5MR, plus some additional country-specific effect, i.e. there is an $f(U_{c,t})$ and a $P_{c,t}$. However, one of the main differences between the two models is that for countries with non-VR data, estimates from the new model can be driven by the data, while the previous model is restricted to follow the trajectory of the U5MR in a particular country, plus or minus some country-specific intercept. The other differences between the two models are highlighted in Table 3.4.

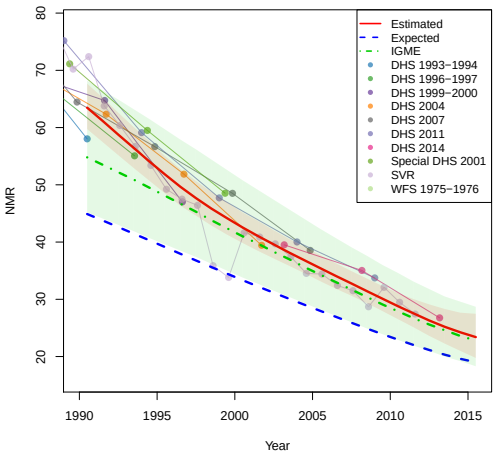
Table 3.4: Comparison of two models

IGME 2014	New model
Model used for non-VR countries	Model used for all countries
Model relation between NMR and U5MR	Model relation between ratio and U5MR
$f(U_{c,t})$ is quadratic	$f(U_{c,t})$ is linear with changing slope
$P_{c,t}$ is a country and region-specific intercept	$P_{c,t}$ is a country-specific intercept + fluctuations
Country-specific effect constant	Country-specific effect can vary
Only considers sampling error	Includes sampling and non-sampling error

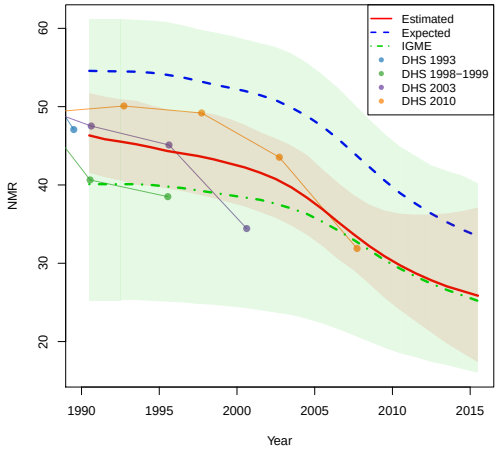
Fig. 3.9 compares the results of four countries to the estimates from the current IGME model. The estimates from the previous IGME model generally follow the same trajectory as the expected line, as determined by U5MR patterns, and is shifted up or down depending on the estimate of the country-specific effect. In contrast, the estimates from the new model follow the data more closely. The fluctuation part of the country-specific multiplier $P_{c,t}$ allows the estimated line to move above or below the expected line, as is the case with the Dominican Republic (Fig. 3.9d). In addition there is generally less uncertainty around the estimates in the new model, especially in periods where there are data.

3.4.5 Model validation

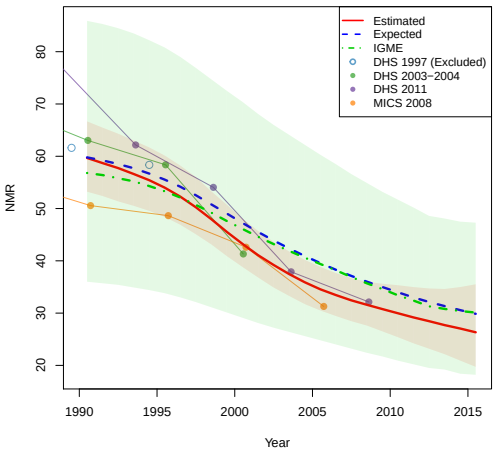
Model performance was assessed through an out-of-sample model validation exercise. In creating a training dataset, rather than removing observations at random, the



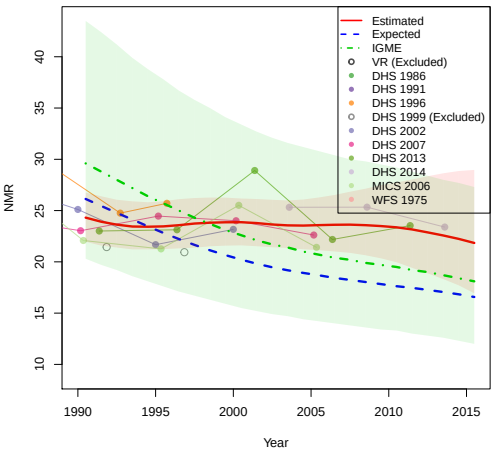
(a) Bangladesh



(b) Burkina Faso



(c) Mozambique



(d) Dominican Republic

Figure 3.9: Estimated neonatal mortality rate(deaths per 1,000 births) for four example countries; new versus IGME 2014 model. The green dashed line and associated shaded area shows the model estimates from the 2014 IGME model.

process of removing data was chosen to emulate the way in which new data may be received Alkema et al. (2012). Mortality databases are updated as least once a year as more data become available. These updates may include not only data for the most recent time period but may also include, for example, retrospective estimates from a survey. Ideally the model should not be sensitive to updates of historical data so estimates do not change from year to year.

The training set was constructed by leaving out the most recent survey data series, and for countries with only one series (including VR countries), the most recent 20% of data observations were removed. The resulting training data set was made of around 80% of the total data available.

For the left-out observations, the absolute relative error is defined by

$$e_i = \frac{|r_{c,i} - \tilde{r}_{c,i}|}{\tilde{r}_{c,i}},$$

where $\tilde{r}_{c,i}$ denotes the posterior median of the predictive distribution for a left-out observation $r_{c,i}$ based on the training set. Coverage is defined by

$$\frac{1}{N}\sum 1[r_{c,i} \geq l_{c[i]}(t[i])]1[r_{c,i} < r_{c[i]}(t[i])]$$

where N is the the total number of left-out observations considered and $l_{c[i]}(t[i])$ and $r_{c[i]}(t[i])$ the lower and upper bounds of the predictions intervals for the ith observation. Coverage at the 80, 90 and 95% levels was considered.

The validation measures were calculated for 100 sets of left-out observations, where each set consisted of a random sample of one left-out observation per country. Table 3.5 shows the median and standard deviation of each validation measure. The median absolute relative error between the observations and estimated value was less than 10%, and the coverage of the prediction intervals is approximately as expected.

Table 3.5: Validation measures, left-out data

	Expected	Median	Std. Dev
Mean absolute relative error	-	0.09	0.02
80% coverage	0.80	0.84	0.02
90% coverage	0.90	0.92	0.02
95% coverage	0.95	0.96	0.02

A similar set of validation measures was calculated comparing the model estimates based on the training data set with the model estimates based on the full data set. The absolute relative error is defined as

$$e_i = \frac{|r^*_{c,i} - \tilde{r}_{c,i}|}{r^*_{c,i}},$$

where $r^*_{c,i}$ and $\tilde{r}_{c,i}$ denote the posterior median of the predictive distribution for the observation $r_{c,i}$ based on estimates using the full and training dataset, respectively. Coverage refers to what proportion of the posterior median estimates from

the training dataset fall within the 80%, 90% and 95% bounds of estimates using the full dataset.

Results in Table 3.6 are reported for estimates up to (and including) 2005, and post 2005. Model performance is better prior to 2005. This is due to the most recent data being removed, so the data prior to 2005 would be very similar between training and test sets. However the post-2005 measures show estimates are reasonably consistent between the reduced and full datasets.

Table 3.6: Validation measures, model comparison

	Expected	≤ 2005	> 2005
Mean absolute relative error	-	0.05	0.09
80% coverage	≥ 0.80	0.90	0.77
90% coverage	≥ 0.90	0.94	0.84
95% coverage	≥ 0.95	0.96	0.90

3.5 Discussion

A new model was introduced for estimating neonatal mortality rates. The model can be expressed as the product of an overall relationship with U5MR and a country-specific effect. The overall relationship with U5MR is a simple linear function, while the country-specific effect is modeled through P-spline regression as a country-specific intercept plus fluctuations around that intercept.

Estimates of the NMR were produced for 195 countries, spanning at least the period 1990–2015. The model appears to perform well in a wide variety of situations where extent and type of data available varies. In many developed countries, where VR data series are complete and uncertainty around the data is low, NMR estimates follow the data closely. On the other hand, where there is limited data available or if uncertainty around the data is high, estimates are more influenced by the trends in U5MR.

The model was fit within a Bayesian hierarchical framework, allowing information about trends in NMR to be exchanged across countries. Through the hierarchical structure, the smoothness in trends in NMR for countries with few data available, or highly uncertain data, is partially informed by countries with more reliable data. While this setup has the potential to introduce biases in country-specific estimates, it allows estimates with reasonable amounts of uncertainty to be produced for all countries, even in the absence of reliable data. Validation exercises did not highlight any problems with bias, suggesting that estimates and uncertainty produced by the model are reliable in a variety of data situations.

Model estimates were compared to estimates from the existing IGME model. The

notable advantage of this model is that trends in NMR for countries without VR data are driven by the data itself, rather than just reflecting trends in U5MR, as is the case with the existing model. Another advantage of this model is that it is along the same methodological lines as the current model used by IGME to estimate U5MR Alkema and New (2014). Estimates produced by this model will help to monitor a country's progress in reducing neonatal mortality and reaching the targets set in the Sustainable Development Goals.

There are several avenues worth investigating in further research. The choice of a linear function with changing slope for $f(U_{c,t})$ was a data-driven decision, based on the observed relationship in Fig. 3.2. It would be interesting to compare the performance of models which have a functional form that draws upon existing demographic models. For example, an extended version of the Brass relational logistic model Brass (1971) and Siler models Siler (1983) can be used to predict survival in the first months of life, as a function of the survivorship at older ages.

The potential for bias in estimates from survey data is always a concern. Bias may occur from interviewing a sample that is not representative of the overall population, from selective omission of answers, and can even be influenced by the length of the survey administered Bradley (2015). The data model included an estimation of an overall level of non-sampling error for each survey type, which may account for some random reporting errors. However, there is scope to further extend the data model to try to better estimate potential bias in survey data estimates.

The focus of this paper was on the methodology. Future work will also focus on interpretation of results. More investigation is needed on what is potentially causing NMR to be higher- or lower-than expected in outlying countries and whether these are real effects or artifacts of data issues. This distinction is an important one and will become even more so as the focus on child mortality continues to shift towards the early months of life.

Chapter 4

Deaths without denominators: using a matched dataset to study mortality patterns in the United States

4.1 Introduction

To understand national trends in mortality over time, it is important to study differences by demographic, socioeconomic and geographic characteristics. For example, the recent stagnation in life expectancy at birth in the United States is largely a consequence of worsening outcomes for males in young-adult age groups (Kochanek et al. (2017)). It is essential to understand differences across groups to better inform and target effective health policies. As such, studying mortality disparities across key subpopulations has become an important area of research. Recent studies in the United States have looked at mortality inequalities across income (Chetty et al. (2016); Currie and Schwandt (2016)), education (Hummer and Lariscy (2011); Masters et al. (2012); Hummer and Hernandez (2013)) and race (Murray et al. (2006); Case and Deaton (2017)), finding evidence for increasing disparities across all groups.

One issue with studying mortality inequalities, particularly by socioeconomic status (SES), is that there are few micro-level data sources available that link an individual's SES with their eventual age and date of death. The National Longitudinal Mortality Study (NLMS) (Sorlie et al. (1995)); National Health Interview Survey Linked Mortality Files (NCHS (2005)); and the Health and Retirement Study (Juster and Suzman (1995)) are important survey-based resources that contain SES, health and mortality information. However, these data sources only contain 10,000-250,000 death records over the period of study, so once the data are disaggregated by year, demographic and SES characteristics, the counts can be quite small and thus uncer-

tainty around mortality estimates is high.

There has been an increasing amount of mortality inequalities research that makes use of large-scale administrative datasets; for example, the use of Social Security (SSA) earnings and mortality data (Waldron (2007)) and income, tax and mortality data from the Internal Revenue Service (IRS) (Chetty et al. (2016)). However, while large in size, these administrative datasets lack richness in terms of the type of information available. The SSA and IRS datasets only include information about income, and not other characteristics such as education or race. In addition, these datasets are not publicly available, which makes validation, reproducibility and extension of the research difficult.

In this paper, a new dataset for studying mortality disparities and changes over time in the United States is presented. The dataset, termed 'CenSoc', uses two large-scale datasets: the full-count 1940 Census to obtain demographic, socioeconomic and geographic information; and that is linked to the Social Security Deaths Masterfile (SSDM) to obtain mortality information. The full-count 1940 census has been used in many areas of demographic, social and economic research since it has been made digitally available (e.g. income inequality (Frydman and Molloy (2011)); education outcomes (Saatcioglu and Rury (2012)); and migrant assimilation (Alexander and Ward (2018))). The SSDM, which contains name, date of birth and date of death information, has been used to study mortality patterns, particularly at older ages (Hill and Rosenwaike (2001); Gavrilov and Gavrilova (2011)). The resulting CenSoc dataset[1] contains over 7.5 million records linking characteristics of males in 1940 with their eventual date of death.

As a consequence of the census and SSDM spanning two separate time points, the mortality information available in CenSoc has left- and right-truncated deaths by age, and no information about the relevant population at risk at any age or cohort. For example, the cohort born in 1910 is observed at age 30 in the 1940 census and has death records for ages 65-95 (observed in the period 1975-2005); however, there is not information on the number of survivors in the same period. Thus, it is not straightforward to use mortality information in CenSoc to create comparable estimates over time. As such, this paper also develops mortality estimation methods to better use the 'deaths without denominators' information contained in CenSoc. Bayesian hierarchical methods are presented to estimate truncated death distributions over age and cohort, allowing for prior information in mortality trends to be incorporated and estimates of life expectancy and associated uncertainty to be produced.

The remainder of the paper is structured as follows. Firstly, the data sources and method used to create the CenSoc dataset are described. The issues with using CenSoc to estimate mortality indicators are then discussed. Two potential methods of mortality estimated are presented: a Gompertz model, and a principal components approach. These models are evaluated based on fitting to United States mortality data available through the Human Mortality Database. The principal components

[1]Version 1 available at: `https://censoc.demog.berkeley.edu/`.

regression framework is then applied to the CenSoc data to estimate mortality trends by education and income. Finally, the results and future work are discussed.

4.2 The CenSoc dataset

The CenSoc dataset was created by combining two separate data sources: the 1940 census, and the Social Security Deaths Master file (SSDM). The two data sources were matched based on unique identifiers of first name, last name and age at the time of the census. Due to issues with potential name changes with marriage, the matching process is restricted to only include males.

As described below, the census observes individuals in 1940, and the SSDM observes individuals in the period 1975-2005. Therefore, by construction, the CenSoc dataset can only contain individuals who died between 1975 and 2005.

4.2.1 Data

The demographic and socioeconomic data come from the U.S. 1940 census, which was completed on 1 April 1940. The census collected demographic information such as age, sex, race, number of children, birthplace, and mother's and father's birthplace. Geographic information, including county and street address, and economic information such as wages, non-wage income, hours worked, labor force status and ownership of house was also collected. The 1940 census had a total of 132,164,569 individuals, 66,093,146 of whom were males.

The 1940 census records were released by the U.S. National Archives on April 2, 2012 (National Archives (2018)). The original 1940 census records were digitized by Ancestry.com and are available through the Minnesota Population Center (MPC). The MPC provides a de-identified version of the complete count census as part of the IPUMS-USA project (Ruggles et al. (2000)). However, names and other identifying information are not available from the IPUMS website. Access to the restricted 1940 census data was granted by agreement between UC Berkeley and MPC. The data are encrypted and can only be accessed through computers or servers on the Berkeley demography network.

Information on the age and date of death was obtained through the SSDM. This contains a record of all deaths that have been reported to the Social Security Administration (SSA) since 1962. The SSDM is used by financial and government agencies to match records and prevent identity fraud and is considered a public document under the Freedom of Information Act. Monthly and weekly updates of the file are sold by the National Technical Information Service of the U.S. Department of Commerce. A copy of the 2011 version was obtained through the Berkeley Library

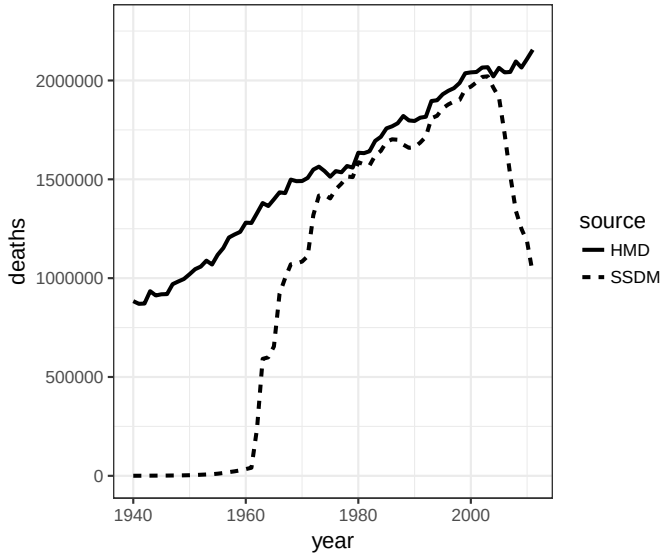


Figure 4.1: Comparison of the number of deaths at ages 55+ in the SSDM (dotted line) and HMD (solid line), 1940-2011. While deaths in the earlier and later periods are underreported in the SSDM, the period 1975-2005 has close to full coverage.

Data Lab.

The SSDM contains an individual's first name, last name, middle initial, social security number, date of birth and date of death. The 2011 file has 85,822,194 death records. The death dates span the years 1962-2011. There are 76,056,377 individuals in the SSDM who were alive at the time of the 1940 census.

Completeness of death reporting in the SSDM is lower pre-1970s, when a substantial proportion of the population did not pay into the social security system. Deaths are more likely to be reported at older ages, when a person was more likely to be receiving social security benefits (Huntington et al. (2013)). Previous studies suggest more than 90% completeness of deaths over age 65 reported in the SSDM since 1975, compared to vital statistics sources (Hill and Rosenwaike (2001)).

To check the coverage of SSDM at the population level, the total number of deaths by year reported in the SSDM was compared to those in the Human Mortality Database (HMD) (HMD (2018)). As Figs. 4.1 and 4.2 illustrate, the completeness of the SSDM file is around 95% for ages 55+ in the period 1975-2005. As such, the data used to create CenSoc is restricted to only include deaths from SSDM that occurred between the period 1975-2005.

4.2.2 Data preparation

For the Census dataset, the following pre-processing steps were done:

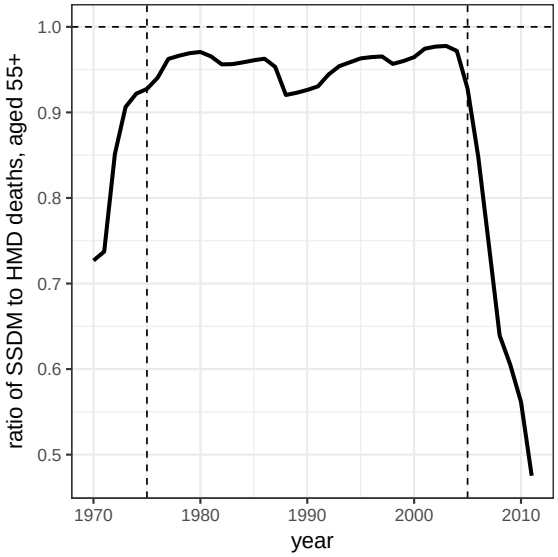


Figure 4.2: Ratio of deaths at ages 55+ in the SSDM to HMD, 1970-2011. The ratio is around 95% over the period 1975-2005.

1. Convert all name strings to upper case.
2. Remove the middle name. The original first name variable contains both first and middle name. The name string is split and only the first name is used for matching.
3. Remove rows where either the first or last name are just question marks or blank.
4. Create a match key by concatenating last name, first name and age.
5. Subset the data to only include males.

For the social security deaths files, there are three raw files in which rows contain a continuous string of characters. For each of the three files, each row is split into social security number, last name, first name, middle initial, date of death and date of birth. The three files are then bound together to create one large file.

The following pre-processing steps are done:

1. Remove any trailing white space from first and last names
2. Split date of birth and date of death to get day, month and year of birth and death.
3. Calculate age of person at census. The age is calculated based on knowing the date of birth and that the 1940 census was run in April.
4. Remove any deaths where the date of birth is missing

5. Remove any deaths of people who born after 1940
6. Remove any deaths before 1975
7. Create a match key by concatenating last name, first name and census age.

4.2.3 Match Method

The two datasets are matched based on exact matches of first name, last name and age. For example, a match key could be `EYREJANE18`. Census records with a key that is not found in the social security deaths database are not matched. The specific steps are:

1. Load in the cleaned census and social security datasets.
2. Remove duplicate keys.
3. Merge the datasets based on key.

Due to file size, the matching step is done separately for each census file in each U.S. state. The resulting national dataset is created by:

1. Loading and binding all state matched files.
2. Removing all rows that have duplicated keys.

4.2.4 Resulting Dataset

A total of 7,564,451 individual males were matched across the census and SSDM to create the CenSoc dataset. As the 1940 full count census had 66,093,146 males, this corresponds to a raw match rate of 11.4%. A total of 43,881,719 males in the census had unique keys; as such the match rate on unique keys was 17.2%.

The raw match rates differ markedly by cohort/age at census. As Table 1 illustrates, match rates are highest for 15-40 year olds. This corresponds to cohorts born in 1900-1925.

Table 4.1: CenSoc match rates by age group. For a particular age group, the % matched is the number of males in CenSoc divided by males in the 1940 census. The % unique matched is the CenSoc number divided by males with a unique key in the 1940 census.

Census age	% matched	% unique matched
0-4	9.1	14.4
5-9	11.6	18.4
10-14	14.5	22.7
15-19	17.0	26.3
20-24	18.2	27.4
25-29	18.0	26.8
30-34	16.6	24.8
35-39	13.9	20.7
40-44	10.8	16.0
45-49	7.4	11.0
50-54	4.2	6.2
55-59	1.7	2.6
60-64	0.4	0.7
65-69	0.1	0.1
70-74	0.0	0.0
75+	0.0	0.0

These raw match rates do not take into consideration mortality. Some individuals died before 1975, and some are still alive after 2005; neither appears in the SSDM. In particular, the low rates at older ages are mostly due to the fact that people of that age in the census have already died by 1975. Thus we would never expect to get match rates of 100% given we only observe a truncated window of deaths.

The matched CenSoc data and unmatched census records were also compared based on a set of socioeconomic variables, to understand the relative representation of key socioeconomic groups. The CenSoc dataset contains a slightly higher proportion of people who completed high school; own their own home; is the household head; living in urban areas; and are white (Fig. 4.3). These differences are relatively small, but consistently show CenSoc contains more advantaged people. There are several potential reasons for this. Firstly, it could be that more advantaged individuals are less likely to die before 1975, and so more likely to be observed in the window of SSDM. Secondly, it could be that more advantaged individuals are more likely to be matched, given they survived to 1975. This could be because they are more likely to have a social security number, and so be included in the dataset, or are less likely to be matched due to data quality issues (nicknames, misspelled names, etc.).

While there are small differences across the matched and unmatched datasets, these are expected given different mortality patterns across subgroups. The somewhat selective nature of the matched CenSoc reocrds means that mortality estimates might

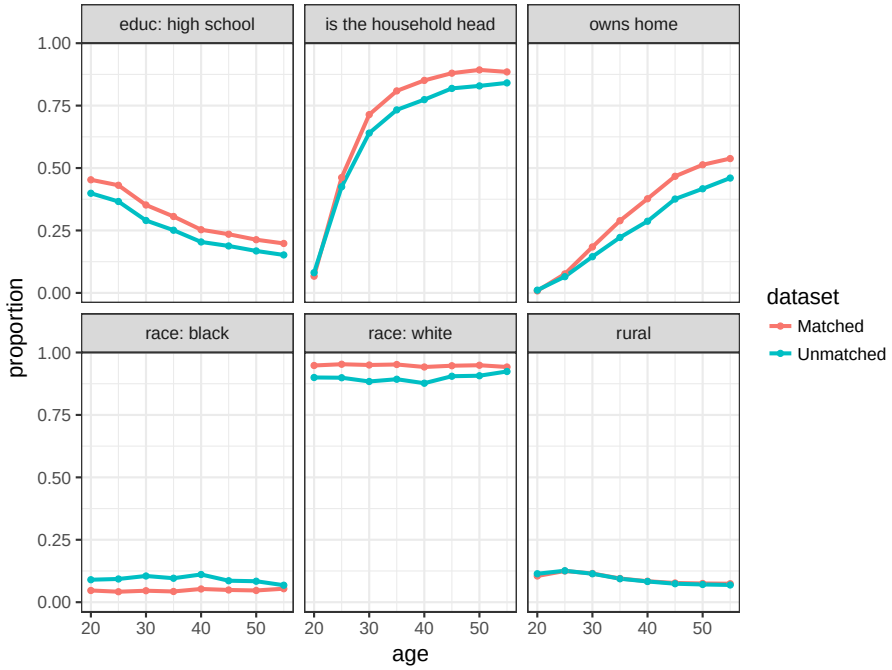


Figure 4.3: Comparison of socioeconomic characteristics of matched and unmatched datasets. The red line is the proportion of each age group with that characteristics in the CenSoc dataset. The blue line is the same proportion for unmatched individuals in the 1940 census.

be slightly lower than in the general population. For the study of subpopulations, however, it matters less if the overall data set is representative by subgroup, as long as the within-group CenSoc sample is broadly representative of that same group in the overall population.

4.3 Issues with using CenSoc to study mortality patterns

The CenSoc dataset contains individual records that link date of birth and death with other demographic and socioeconomic information. it is a useful resource to study patterns in mortality inequalities over time. However, as a consequence of the CenSoc data being constructed from two different data sources available in different years, it is not necessarily straightforward to calculate unbiased estimates of mortality differences by subgroup and over time. This section describes the issue and motivates the methods for mortality estimation described in later sections.

4.3.1 If complete death records were available

Instead of the CenSoc dataset, imagine if we could track every person in the 1940 census until they died, so the available dataset contained full records of death for all persons. If this were the case, then we could use standard demographic and survival analysis approaches to calculate key mortality indicators by subpopulation.

If perfect deaths data existed, we would have a complete record of the number people by cohort (who were alive in the 1940 census) and the ages at which they died. For extinct cohorts, these death counts could be used to construct cohort lifetables and mortality indicators such as life expectancy or variance in age of death could be compared. Lifetables could also be constructed by key socioeconomic groups such as income, race or education, and change in mortality tracked over cohort.

For cohorts that are not yet extinct, these data would be right-censored, i.e. the last date of observation (in this fictitious dataset, this would be 2018) is before the observed time of death. However, standard techniques from survival analysis could be used to measure mortality indicators. For example, non-parametric techniques like the Kaplan-Meier estimator could be used to compare empirical survival curves, and differences in survival across groups could be estimated in multivariate setting using Cox proportional hazard models (Hougaard (2012); Wachter (2014)).

4.3.2 Characteristics of CenSoc data

However, we do not have complete records of dates of death for all persons in the 1940 Census. Instead, after observing the full population in 1940 (the blue line in the Lexis diagram, Fig. 4.4), we observe deaths only over the period 1975-2005 (the red shaded area in Fig. 4.4). This creates issues for the estimation of mortality indicators, for two main reasons: firstly, the deaths data available is both left- and right-truncated, and secondly, we do not observe the population at risk of dying at certain ages.

As deaths are only observed over the period 1975-2005, the number of people who died before 1975, and the number who are still yet to die after 2005, are unknown. For the older cohorts, many have already died before 1975. The younger cohorts, e.g. those born in 1940, will only have reached relatively young ages by 1975, so many are still yet to die.

Fig. 4.5 illustrates the left- and right-truncation in the CenSoc dataset. Each colored line is a different cohort. For each cohort a different set of ages is available; for example, for the 1920 cohort we observe deaths from age 55. In contrast, for the 1890 cohort we only observe deaths from age 85. Thus, methods of mortality estimation need to take the differing truncation into account, and adjust accordingly to make measures comparable over time.

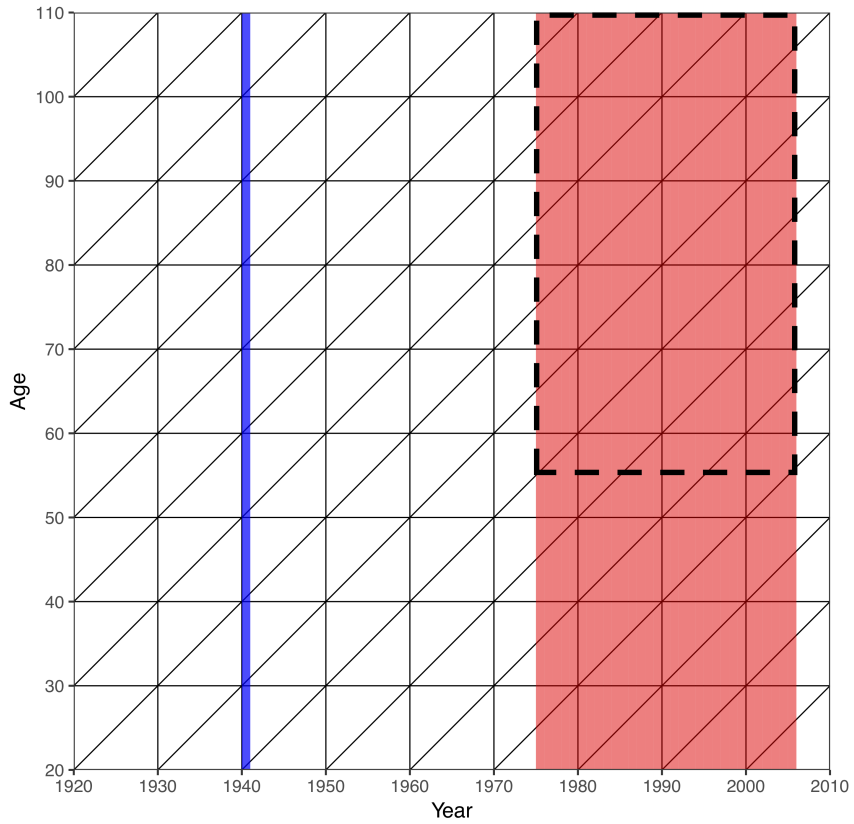


Figure 4.4: Available information for CenSoc. Socioeconomic information is observed in the 1940 census (blue line); deaths are observed over 1975-2005 in the SSDM (red area). We only consider deaths above age 55, indicated by the dashed box.

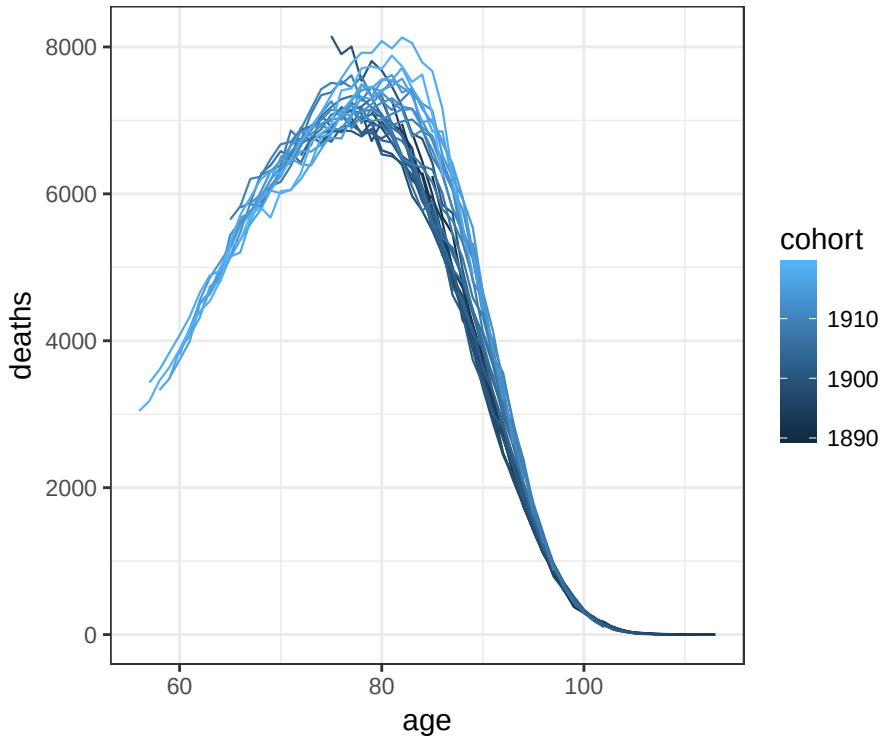


Figure 4.5: Number of deaths observed by age and cohort in CenSoc. Each line is a different cohort. Every cohort has a different set of observed ages available.

While truncation of observed deaths makes mortality estimation across cohorts more difficult, it is still possible with existing techniques. For example, techniques such as Kaplan-Meier and Cox proportional hazards regression are still possible with truncated and censored observations (Hougaard (2012)). If parametric models are used, truncation can be incorporated into the death density and survivorship functions (Nelson (2005)). However, it is the combination of truncated observations with the fact that no denominators are observed that makes estimation more difficult.

In the period 1975-2005, we only observe deaths, not the total population. Not all persons in the 1940 Census are matched in CenSoc. There is no way of knowing whether unmatched people were still alive in 1975 or not. Therefore, we do not know the size of the population at risk of dying in 1975.

Knowing the exposure to risk is important for most mortality indicators. Lifetable quantities such as survivorship, the probability of dying and the hazard rate all rely on calculating some measure of risk relative to the baseline population.

For extinct cohorts, we can assume there are no survivors beyond the ages that we observe and so it is possible to use the reverse survival method (Andreev et al. (2003)) or multivariate techniques such as Cox proportional hazards regression to study differences in survival across socioeconomic groups. However, for cohorts that are not extinct, coefficient estimates from Cox regression will be biased towards zero. Thus, other techniques of mortality estimation need to be developed.

4.4 Mortality estimation for data with no denominators

In this section, methods of mortality estimation for use with CenSoc are introduced. Firstly, relevant survival quantities are defined. The focus is then on estimating the distribution of deaths by age, which is the relevant quantity for CenSoc. Two models for deaths distribution estimation are introduced, one parametric (Gompertz) and one semi-parametric (principal components). The models are described and relative performance is assessed based on fitting to U.S. mortality data available through the Human Mortality Database.

4.4.1 Definition of survival quantities

Define the survivorship function as

$$l(x) = Pr(X > x) \tag{4.1}$$

i.e. $l(x)$ is the probability that the age of death, X, is greater than x, or in other words, the proportion of the population that survive to exactly age x. The hazard function is

$$\mu(x) = \lim_{\Delta x \to 0} \frac{Pr(x \leq X < x + \Delta x | x \leq X)}{\Delta X}. \tag{4.2}$$

This is equivalent to

$$\mu(x) = \frac{-d \log l(x)}{dx}. \tag{4.3}$$

The cumulative death distribution function is

$$D(x) = 1 - l(x) \tag{4.4}$$

and the density function is the derivative of this, i.e.

$$d(x) = \frac{-dl(x)}{dx} = \mu(x)l(x) = \mu(x) \exp(-M(x)) \tag{4.5}$$

where $M(x) = \int_0^x \mu(u)du$. As all people in a population must die eventually, *memento mori*, $d(x)$ is a probability density function, with $\int d(x) = 1$. These quantities are continuous across age. Throughout this paper, the discrete versions are denoted with a subscript x, for example, the discrete death distribution is written as d_x.

Given the lack of denominators in the CenSoc data set, the focus is on estimating mortality across cohorts based on information available about the density of deaths, d_x. As shown in Fig. 4.5, we have partial information about the shape of d_x across cohorts. As such, the estimation of d_x, i.e. the discrete death distribution by single year of age, is a starting point for inference about other mortality quantities. Once

we have information about d_x, other lifetable values can be calculated (Wachter (2014), see Section 7.3).

Consider the following to illustrate how the estimation of d_x relates to the observed death counts. Say we observe death counts by age, y_x, which implies a total number of deaths of D, i.e.

$$\sum_x y_x = D.$$

If we multiply the total number of deaths D by d_x, then that gives the number of deaths at age x. In terms of fitting a model, we want to find estimates of the density, d_x, which best describes the data we observe, y_x.

4.4.2 Accounting for truncation

Eq. 4.5 gives the density of deaths over the entire age range x. Suppose instead we only observe ages between x_L and x_U. In order to remain a probability density function, $d(x)$ for the truncated period, defined as $d^*(x)$, needs to be divided through by the difference of the survivorship functions at each end point:

$$d^*(x) = \frac{d(x)}{l(x_L) - l(x_U)}. \tag{4.6}$$

4.4.3 Estimating the death distribution

Define y_x to be the observed number of deaths at age x. It is assumed that deaths are only observed in the window of ages $[x_L, x_U]$. Conditional on the number of the total number of deaths, N, the observed sequence of deaths $\mathbf{y} = y_1, y_2, \ldots, y_n$ has a multinomial distribution (Chiang (1960)):

$$\mathbf{y}|N \sim Multinomial(N, \mathbf{d}^*) \tag{4.7}$$

where $\mathbf{d}^* = d_1^*, d_2^*, \ldots, d_n^*$ and d_x^* is the discrete version of the truncated deaths density, and is equal to the proportion of total deaths that are observed between ages x and $x+1$. The total number of observed deaths, $D = \sum_x y_x$ is Poisson distributed around the true number of deaths i.e.:

$$D \sim Poisson(N). \tag{4.8}$$

Thus, the marginal distribution of y_x is also Poisson distributed (McCullagh and Nelder (1989)), with

$$y_x \sim Poisson(\lambda_x) \tag{4.9}$$

where $\lambda_x = N \cdot d_x^*$. The likelihood function of an observed sequence of deaths $\mathbf{y} = y_1, y_2, \ldots, y_n$ can then be written as:

$$P(\mathbf{y}|\lambda(\theta)) = \exp\left(-\sum_i \lambda_i\right) \frac{\prod_i \lambda_i^{y_i}}{\prod_i y_i!} \tag{4.10}$$

with corresponding log-likelihood

$$l(\mathbf{y}|\lambda(\theta)) = -\sum_i \lambda_i + \sum_i y_i \lambda_i - \log \prod_i y_i!. \tag{4.11}$$

Here, θ refers to the (potentially multiple) parameters that govern the rate of deaths, λ_x. In practice it is the parameters θ we are trying to estimate.

One option to find values of θ is to use maximum likelihood (ML) estimation. In this approach, the true parameter values are assumed to be fixed, but unknown. ML estimation finds parameter values that maximize the log-likelihood function based on data we observe about counts of death by age. Given the complexity of the likelihood function, numerical techniques need to be used, such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Fletcher (2013)).[2] Standard errors around the estimates can be calculated based on the Hessian matrix, and inference can be carried out based on the assumption that the sampling distribution of the parameters are asymptotically normal.

An alternative strategy to find best estimates of θ is to use Bayesian analysis. In contrast to ML estimation, Bayesian methods assume the parameters θ themselves are random variables. The goal is to estimate the posterior distribution of the parameters, $P(\lambda(\theta)|y)$. By Bayes Rule,

$$P(\lambda(\theta)|y) = \frac{P(y|\lambda(\theta)) \cdot P(\lambda(\theta))}{P(y)} \tag{4.12}$$

where $P(y|\lambda(\theta))$ is the likelihood function, $P(\lambda(\theta))$ is the prior distribution on the parameters of interest and $P(y)$ is the marginal probability of the data.

For some posterior distributions, integrals for summarizing posterior distributions have closed-form solutions, or they can be easily computed using numerical methods. However, in many cases, the posterior distribution is difficult to handle in closed form. In such cases, Markov Chain Monte Carlo (MCMC) algorithms can be implemented to sample from the posterior distribution. For example, the Gibbs sampling algorithm (Gelfand and Smith (1990)) generates an instance from the distribution of each parameter in turn, conditional on the current values of the other parameters. It can be shown that the sequence of samples constitutes a Markov chain, and the stationary distribution of that Markov chain is the sought-after joint distribution. Gibbs Sampling can be implemented in R using the JAGS software (Plummer (2012)).

There are several benefits of the Bayesian approach. Firstly, Bayesian methods are generally more computationally efficient than ML approaches, which can be sensitive to initial conditions and can take a relatively long time to converge. Secondly, if we can summarize the entire posterior distribution for a parameter, there is no

[2]The BFGS algorithm, which is a class of quasi-Newton optimization routines, can be implemented using the 'optim' function in R.

need to rely on asymptotic arguments about the normality of the distribution. Having the entire posterior distribution for a parameter allows for additional tests and summaries that cannot be performed under a classical likelihood-based approach. Uncertainty intervals around parameter estimates can easily be calculated through assessing the quantiles of the resulting posterior distribution. In addition, distributions for the parameters in the model can be easily transformed into distributions of quantities that may be of interest are not directly estimated as part of the model. This is especially important in this context, because we are estimating parameters θ, but would like to also calculate implied quantities such as hazard rates or life expectancy.

Another important aspect of the Bayesian approach is that it allows prior information about the parameters to be incorporated into the model. For example, it is expected that the mode age at death of the deaths distribution should be in the range of 70-85, and generally increase over time. Informative priors can be included in the model to incorporate this information. Thus, given these advantages, the methods proposed in the following sections will be fit within a Bayesian hierarchical framework.

4.5 Truncated Gompertz approach

The first approach to estimate the truncated deaths distribution d_x^* is the Gompertz model (Gompertz (1825)). This model is one of the most well-known parametric mortality models. It does remarkably well at explaining mortality rates at adult ages across a wide range of populations, with just two parameters. The Gompertz hazard at age x, $\mu(x)$, has the exponential form

$$\mu(x) = \alpha e^{\beta x}. \tag{4.13}$$

The α parameter captures the starting level of mortality and the β parameter gives the rate of mortality increase over age. On the log scale, Gompertz hazards are linearly increasing across age:

$$\log \mu(x) = \alpha + \beta x \tag{4.14}$$

Note here that x refers to the starting age of analysis and not necessarily age = 0. Indeed, in practice, the assumption of constant log-hazards is not realistic in younger age groups. In this application we are interested in modeling adult mortality, so younger ages are not an issue. There is, however, some evidence of mortality deceleration in the older ages (ages 90+), which would also lead to non-Gompertzian hazards (Kannisto (1988); Horiuchi and Wilmoth (1998)). Other parametric models have been proposed to account for this deceleration, which commonly include additional terms as well as the Gompertz α and β (Feehan (2017)). The most parsimonious parametric approach is illustrated; however it could be extended to models with more parameters.

Given the relationship between the hazard function and the survivorship function given in Eq. 4.3, the expression for the Gompertzian survivorship function is

$$l(x) = \exp\left(-\frac{\alpha}{\beta}\left(\exp(\beta x) - 1\right)\right) \tag{4.15}$$

and it follows from Eq. 4.5 that the density of deaths at age x, $d(x)$ is

$$d(x) = \mu(x)l(x) = \alpha \exp(\beta x) \exp\left(-\frac{\alpha}{\beta}\left(\exp(\beta x) - 1\right)\right). \tag{4.16}$$

4.5.1 Reparameterization

Estimates of the level and slope parameters α and β in the Gompertz model are highly correlated. In general, the smaller the value of β, the larger the value of α (Tai and Noymer (2017)). For example, Fig. 4.6 shows values of α and β that lead to mode ages of death within a plausible range (see Eq. 4.17 below). The figure illustrates two main points. Firstly, the plausible values of α and β for human populations fall within a relatively small interval: α is not likely to be greater than 0.006, and β is not likely to be greater than 0.15. Secondly, the strong negative correlation between the two parameters is apparent. A simulated study showed the correlation between estimated values of α and β can be upwards of 0.95 (Missov et al. (2015)), which is a statistical artifact rather than giving any insight into the ageing process or heterogeneity in frailty (Burger and Missov (2016)).

The correlation between these parameters can cause estimation issues. As such, following past research (Missov et al. (2015); Vaupel and Missov (2014)) a re-parameterized version of the Gompertz model in terms of the mode age is considered. Under a Gompertz model, the mode age at death, M is (Wachter (2014))

$$M = \frac{1}{\beta}\log\left(\frac{\beta}{\alpha}\right). \tag{4.17}$$

Gompertz hazards can thus be reparameterized in terms of M and β:

$$\mu(x) = \beta \exp\left(\beta(x - M)\right). \tag{4.18}$$

As Missov et al. (2015) note, M and β are less correlated than α and β. In addition, the modal age has a more intuitive interpretation than α. The expression for the truncated deaths density d_x^* follows in the same way from Eqs. 4.5 and 4.6:

$$\begin{aligned} d^*(x) &= \frac{\mu(x) \cdot l(x)}{l(x_L) - l(x_U)} \\ &= \frac{\beta \exp\left(\beta(x - M)\right) \cdot \exp\left(-\exp\left(-\beta M\right)\left(\exp(\beta x) - 1\right)\right)}{\exp\left(-\exp\left(-\beta M\right)\left(\exp(\beta x_L) - 1\right)\right) - \exp\left(-\exp\left(-\beta M\right)\left(\exp(\beta x_U) - 1\right)\right)} \end{aligned} \tag{4.19}$$

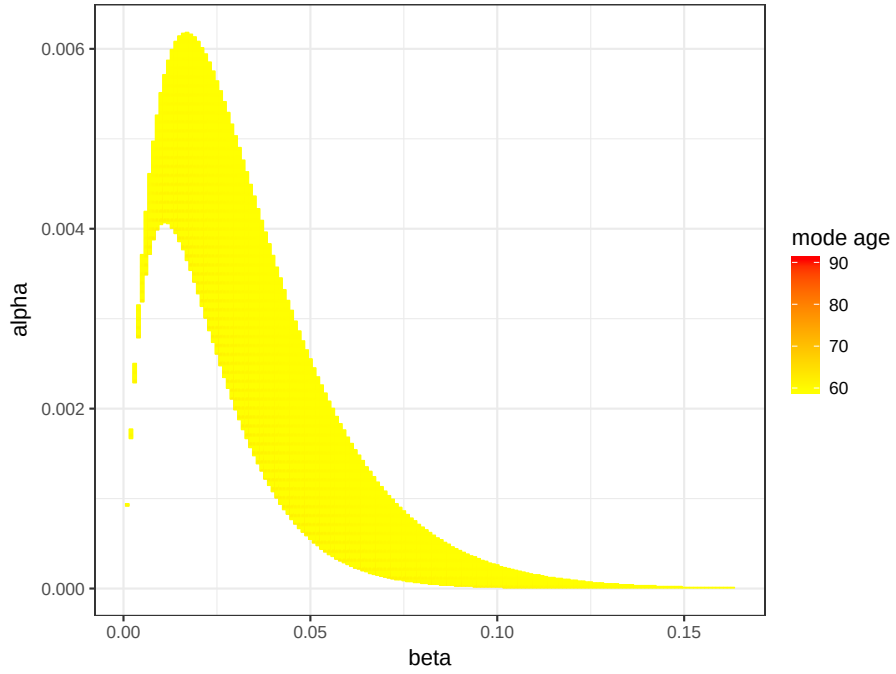


Figure 4.6: Plausible values of Gompertz parameters α and β given a mode age of between 60-90 years. In general, the larger the value of α, the smaller the value of β. The values of α and β are limited to be below 0.006 and 0.15, respectively.

4.5.2 Bayesian hierarchical model

Eq. 4.20 gives a parametric expression for the distribution of deaths between ages x_L and x_U in terms of two parameters, β and M. This section describes a strategy to estimate these parameters and associated uncertainty.

Often when fitting a Gompertz process to observed mortality data, estimates of α and β are obtained by regression techniques of mortality rates by age, based on Eq. 4.14. For example, a recent paper by Tai and Noymer compared different the performance of difference regression techniques in fitting Gompertz models to data from the Human Mortality Database (HMD) (Tai and Noymer (2017)). However, in this situation, as discussed in Section 4.4.3, parameters need to be estimated based on the non-linear deaths density d_x^*.

We propose a Bayesian hierarchical framework to estimate β and M over cohorts. Firstly, assume that we observe counts by age and cohort $y_{c,x}$ between the ages $[x_{c,L}, x_{c,U}]$. Note the truncated age window can vary by cohort. The total number of deaths observed by cohort is equal to D_c.

From Eqs. 4.9 and 4.8, the observed deaths by age and cohort are distributed

$$D_c \sim Poisson(N_c) \tag{4.20}$$
$$y_{c,x} \sim Poisson(\lambda_{c,x}) \tag{4.21}$$

where $\lambda_{c,x} = N_c \cdot d^*_{c,x}$. In words, the total number of observed deaths in a cohort are a realization of a Poisson process with rate N_c. The observed death counts by age are a realization of a Poisson process with a rate equal to the total deaths multiplied by the proportion of total deaths occurring at that age. In the Gompertz set up, from Eq. 4.20 we have

$$d^*_{c,x} = \frac{\mu(c,x) \cdot l(c,x)}{l(c,x_L) - l(c,x_U)}$$

where $\mu(c,x) = \beta \exp\left(\beta_c(x - M_c)\right)$ and $l(c,x) = \exp\left(-\exp\left(-\beta_c M_c\right)\left(\exp(\beta_c x) - 1\right)\right)$.

Priors on M_c and β_c

As part of the framework, prior distributions need to be specified on the M_c and β_c parameters. One option would be to put uninformative priors on both parameters, which treat each cohort independently. For example, relatively uninformative priors would be

$$M_c \sim U(50, 90)$$

and

$$\beta_c \sim U(0.0001, 0.2).$$

That is, both parameters are draws from Uniform distributions with bounds determined by plausible values of mortality (Fig. 4.6). However, this is modeling each cohort separately and does not allow for cohorts that may have fewer observed ages of death available to be informed by estimates of past cohorts. The value for β could increase or decrease over time, depending on the balance of mortality shifting and mortality compression (Tuljapurkar and Edwards (2011); Bergeron-Boucher et al. (2015); Tai and Noymer (2017)). However, we know from past trends that the mode age at death has been increasing fairly steadily across cohorts in developed countries (Paccaud et al. (1998); Wilmoth and Horiuchi (1999); Canudas-Romo (2008)). Thus we could incorporate this knowledge into the model in the form of a prior on M_c that has a temporal structure. For example, we chose to model M_c as a second-order random walk:

$$M_c \sim N(2M_{c-1} - M_{c-2}, \sigma^2_M).$$

This set-up penalizes deviations away from a linear trend, and so the fit of M_c, especially over shorter time periods, is relatively linear. Second-order random walk priors have been used in past mortality modeling approaches (e.g Alkema and New (2014); Currie et al. (2004)). Other prior options for M_c could include a linear model over cohort, or a times series model with drift; however the second-order random

walk is less restrictive. The full model set-up becomes:

$$
\begin{aligned}
D_c &\sim Poisson(N_c) \\
y_{c,x} &\sim Poisson(\lambda_{c,x}) \\
\lambda_{c,x} &= N_c \cdot d^*_{c,x} \\
d^*_{c,x} &= \frac{\beta_c \exp\left(\beta_c(x - M_c)\right) \cdot \exp\left(-\exp\left(-\beta_c M_c\right)\left(\exp(\beta_c x) - 1\right)\right)}{\exp\left(-\exp\left(-\beta_c M_c\right)\left(\exp(\beta_c x_{c,L}) - 1\right)\right) - \exp\left(-\exp\left(-\beta_c M_c\right)\left(\exp(\beta_c x_{c,U}) - 1\right)\right)} \\
\beta_c &\sim U(0.0001, 0.2) \\
M_c &\sim N(M_{c-1} - M_{c-2}, \sigma^2_M) \\
\sigma_M &\sim U(0, 40)
\end{aligned}
$$

4.6 Principal components regression approach

The Gompertz model relies on two parameters, which, while providing model parsimony, means the shape of the Gompertz death distribution is quite inflexible and may not be able to pick up real patterns in the observed data. There are many other parametric mortality models that could be considered, which include additional parameters for increased flexibility. For example, the Gompertz-Makeham model includes an additional parameter that is age-independent and aims to capture background/extrinsic mortality (Makeham (1860)). The Log-Quadratic model (Steinsaltz and Wachter (2006); Wilmoth et al. (2012)) includes an additional parameter again to account for deceleration of mortality at advanced ages.

While increasing the number of parameters in models increases the flexibility of the fit, this increased complexity means models are also often more difficult to fit, and there may be identifiable issues with some parameters (Willemse and Kaas (2007); Girosi and King (2008)). In addition, increasing the number of parameters may lead to model over-fitting.

As an alternative to more complex parametric models, this section proposes a model framework based on data-derived principal components. The main idea is to use information about underlying mortality trends from existing data sources (a 'mortality standard') to form the basis of a mortality model. Main patterns in death distributions from data are captured via Singular Value Decomposition (SVD) of age-specific death distributions. The SVD extracts 'principal components', which describe main features of death distributions.

Principal components create an underlying structure of the model in which the regularities in age patterns of human mortality can be expressed. These can be used as a basis for a regression framework to fit to the dataset of interest. Thus, instead of modeling d^*_x as a parametric distribution, as in Eq. 4.20, the model for d^*_x will be based on a principal components regression:

$$\text{logit } d^*_x = P_{0,x} + \beta_1 P_{1,x} + \beta_2 P_{2,x} \tag{4.22}$$

where

- $P_{0,x}$ is the mean death distribution (on the logit scale), derived from a mortality standard;
- $P_{1,x}$ and $P_{1,x}$ are the first two principal components derived from the de-meaned mortality standard; and
- The β_d's are the coefficients associated with the principal components.

Many different kinds of shapes of mortality curves can be expressed with different plausible values of the β's. The death distribution is modeled on the logit scale and then transformed after estimation to ensure the estimated values are between zero and one.

The use of SVD in demographic modeling and forecasting gained popularity after Lee and Carter used the technique as a basis for forecasting U.S. mortality rates (Lee and Carter (1992)). More recently, SVD has become increasingly used in demographic modeling, in both fertility and mortality settings. Girosi and King (2008) used this approach to forecast cause-specific mortality. Schmertmann et al. (2014) used principal components based on data from the Human Fertility Database to construct informative priors to forecast cohort fertility rates. Clark (2016) use SVD as a basis for constructing model lifetables for use in data-sparse situations. Alexander et al. (2017) used principal components to estimate and project subnational mortality rates. The SVD/principal components approach seems particularly suited to many demographic applications, due to the nature of demographic indicators being fairly stable across age and changing relatively gradually over time.

4.6.1 Obtaining principal components

The SVD of matrix X is

$$X = UDV^T.$$

The three matrices resulting from the decomposition have special properties:

- The columns of U and V are orthonormal, i.e. they are orthogonal to each other and unit vectors. These are called the left and right singular vectors, respectively.
- D is a diagonal matrix with positive real entries.

In practice, the components obtained from SVD help to summarize some characteristics of the matrix that we are interested in, X. In particular, the first right singular

vector (i.e. the first column of V) gives the direction of the maximum variation of the data contained in X. The second right singular vector, which is orthogonal to the first, gives the direction of the second-most variation of the data, and so on. The U and D elements represent additional rotation and scaling transformations to get back the original data in X.

SVD is useful as a dimensionality reduction technique: it allows us to describe our dataset using fewer dimensions than implied by the original data. For example, often a large majority of variation in the data is captured by the direction of the first singular vector, and so even just looking at this dimension can capture key patterns in the data. SVD is closely related to Principal Components Analysis: principal components are derived by projecting data X onto principal axes, which are the right singular vectors V.

The mortality standard: non-U.S. HMD data

To build a principal components regression framework, we need to choose a suitable mortality 'standard', which forms the basis of the age-specific matrix of death distributions on which the SVD is performed. The chosen standard is based on cohort mortality information available through the HMD, excluding data for the U.S.. In this way, we obtain information about mortality patterns using all available high-quality data, without twice-using the U.S. data. This will enable the validation of models without overfitting. The proportion of total deaths between ages 50-105 at each age was calculated for each available cohort and country. This was done by multiplying the death rates and exposure to get an implied number of deaths by age, then calculating each age as a proportion of total deaths. The cohorts and countries used in the standard were restricted to those that have full information available across all ages.

Fig. 4.7 shows the HMD data on death distributions by cohort and country from which the principal components are derived. Note that the distributions are plotted on the logit scale. Data are available from 23 different countries, across 118 different cohorts from 1850-1910. For some countries and cohorts, the death proportions are quite noisy, for example for many of the cohorts in Israel (ISL). However, the idea of SVD is that the first few principal components will pick up the main patterns in these death distributions.

SVD is performed on a matrix of demeaned logit proportions of deaths at each age between 50 and 105. The matrix has dimensions of 1129×56, as there are 1129 country-cohort observations and 56 ages. Fig. 4.8 shows the mean death distribution and first two principal components obtained from this matrix.[3] The mean schedule gives a shape of baseline mortality across the ages, with mortality peaking

[3]Note we refer to the right singular vectors as 'principal components'. They are technically 'principal axes'.

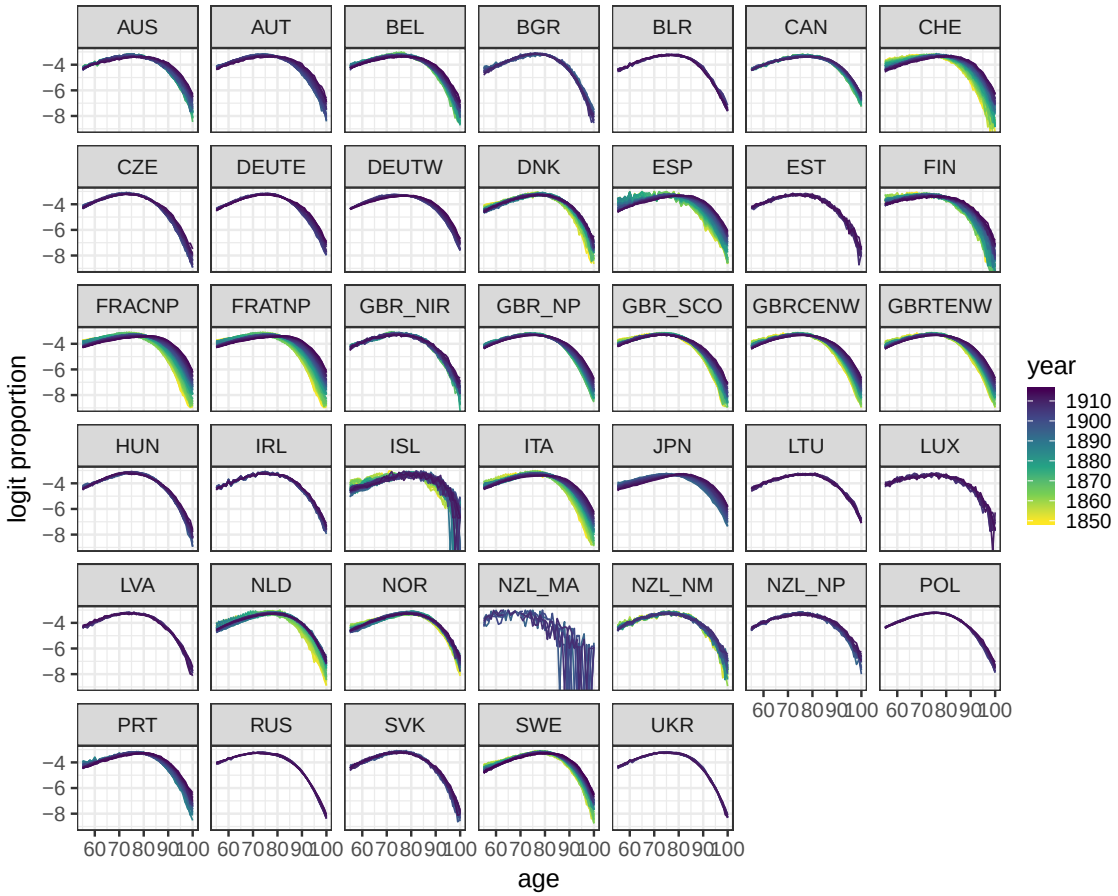


Figure 4.7: Death distributions in HMD by country and cohort, ages 55-105. The plots show the proportion of deaths at each age, plotted on the logit scale. Each line is a different cohort.

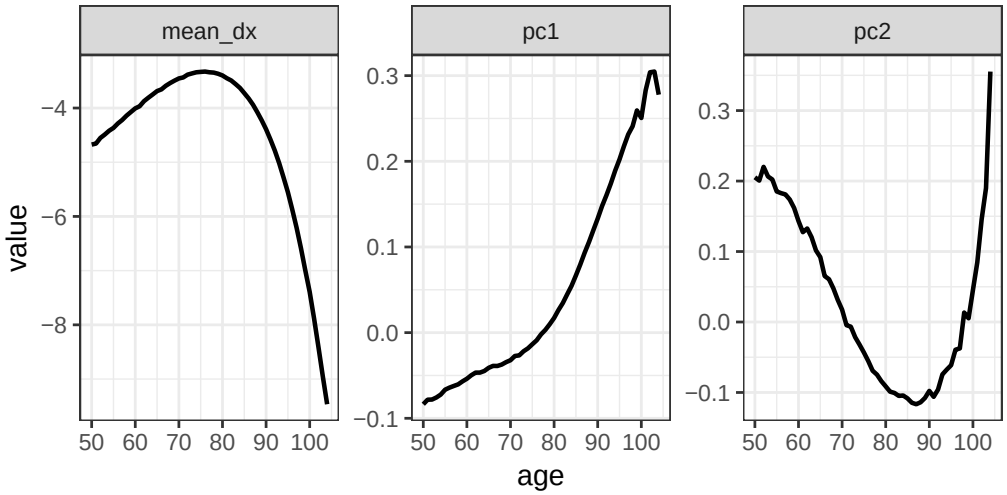


Figure 4.8: Mean death schedule and first two principal components derived from HMD data shown in Fig. 4.7. The components are derived from data transformed to be on the logit scale.

at around age 75. The first principal component could be interpreted as the average contribution of each age to mortality change over time. Note that there is a sign switch of this component at around age 80: proportions at younger ages decrease over time, whereas proportions at older ages increase. The second principal component is related to the shift or compression of mortality around the mode age at death over time.

To reiterate, the idea is to use these three components as the basis of a regression framework. Different values of the regression coefficients lead to different death distributions. Fig. 4.9 shows two example death distributions that can be derived from the combination of the curves shown in in Fig. 4.8. For the red curve, the coefficient on the first principal component is relatively low, and the coefficient on the second component is relatively high, meaning that deaths are shifted to the left and more spread out compared to the blue curve.

4.6.2 Bayesian hierarchical model

The three principal components described above are used as the basis of a regression model within a hierarchical framework to model death distributions over cohorts.

As before we have observed counts by age and cohort $y_{c,x}$ between the ages $[x_{c,L}, x_{c,U}]$. The sum of these observed deaths is equal to D_c. As before we have

$$\begin{aligned} D_c &\sim Poisson(N_c) \\ y_{c,x} &\sim Poisson(\lambda_{c,x}) \end{aligned}$$

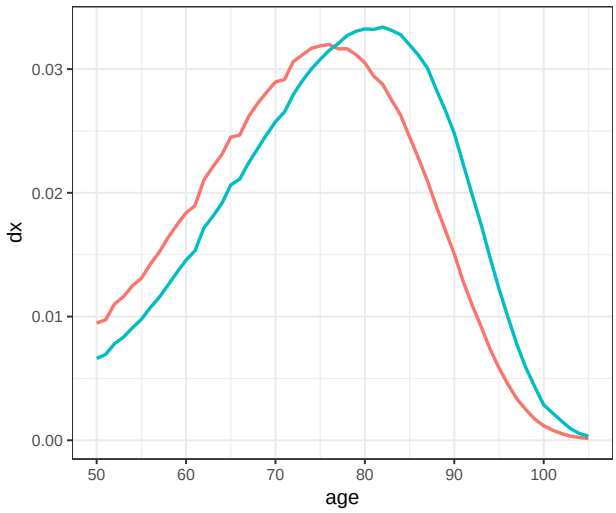


Figure 4.9: Two example death distributions based on different linear combinations of curves in Fig. 4.8. For the red curve, the equation is logit $^{-1}(P_{0,x} + 2.5P_{1,x} + 1.25P_{2,x})$. For the blue curve the equation is logit $^{-1}(P_{0,x} + 6P_{1,x} + 0.8P_{2,x})$.

where $\lambda_{c,x} = N_c \cdot d^*_{c,x}$. Now $d^*_{c,x}$ is modeled on the logit scale as

$$\text{logit } d^*_{c,x} = P_{0,x} + \beta_{1,c}P_{1,x} + \beta_{2,c}P_{2,x} \tag{4.23}$$

where

- $P_{0,x}$ is the mean death distribution on the logit scale.
- The $P_{1,x}$ and $P_{2,x}$ are the first two principal component of the standard logit death distributions, shown in the second and third panels of Fig. 4.8.
- The $\beta_{d,c}$s are the coefficients associated with the principal components.

Note that this is a two parameter model for each cohort, with each of the $\beta_{d,c}$ needing to be estimated. In a similar way to the Gompertz model, each cohort could be modeled independently, with non-informative priors put on the β coefficients. However, estimates of β are likely to be autocorrelated, and so a time series model is placed on the $\beta_{d,c}$'s. Assuming a temporal model on the principal component coefficients aids in the sharing of information about mortality distributions across cohorts, allowing cohorts with relatively few available data points to be partially informed by more data-rich cohorts.

Looking at the interpretation of the principal components used in the model (Fig. 4.8), the first principal component most likely represents a shift in mortality away from younger ages and towards older ages. As such, we expect the coefficient on this principal component to broadly increase over time. As such, similarly to the model age parameter in the Gompertz model, the second-order differences in the $\beta_{1,c}$ are

penalized, which is equivalent to penalizing fluctuations away from a linear trend, while still allowing for a certain degree of flexibility in the trend over time.

$$\beta_{1,c} \sim N(2\beta_{1,c-1} - \beta_{1,c-1}, \sigma_1^2).$$

In terms of principal component 2, it is less clear intuitively what the trends should be over time. As such coefficients are modeled as a random walk across cohorts, which is slightly less restrictive than the model for $\beta_{1,c}$:

$$\beta_{2,c} \sim N(\beta_{2,c-1}, \sigma_2^2)$$

Constraint on d_x^*

For the principal components regression model, there needs to be an additional constraint placed of the principal components $\beta_{d,c}$. The model as explained above does not necessarily ensure that the sum of the resulting deaths distribution d_x^* equals 1. However, this is a fundamental property of d_x^*: over the population of interest, all people must die eventually. As such, an additional constraint is added to the model to ensure that $\sum d_x^* = 1$.

By imposing $\sum d_x^* = 1$, combinations $\beta_{d,c}$ that lead to the constraint not being met are given a probability of 0. In practice, initial values of $\beta_{d,c}$ need to be specified in order to ensure the Gibbs Sampler stays within the constraint. To obtain plausible initial values, the model was first run with no constraint, and then initial values were chosen based on the unconstrained estimates which were close to resulting in $\sum d_x^* = 1$.

The full principal components model set-up is:

$$\begin{aligned}
D_c &\sim Poisson(N_c) \\
y_{c,x} &\sim Poisson(\lambda_{c,x}) \\
\lambda_{c,x} &= N_c \cdot d_{c,x}^* \\
\text{logit } d_{c,x}^* &= P_{0,x} + \beta_{1,c}P_{1,x} + \beta_{2,c}P_{2,x} \\
d_c^* &= \sum_x d_{c,x}^* = 1 \\
\beta_{d,c} &\sim N(2\beta_{d,c-1} - \beta_{d,c-2}, \sigma_d^2) \text{ for } d = 1 \\
\beta_{d,c} &\sim N(\beta_{d,c-1}, \sigma_d^2) \text{ for } d = 2 \\
\sigma_d &\sim U(0, 40)
\end{aligned}$$

4.7 Illustration and comparison of models

The performance of the two models is illustrated by fitting to U.S. mortality data obtained through the HMD (HMD (2018)). This section describes the data avail-

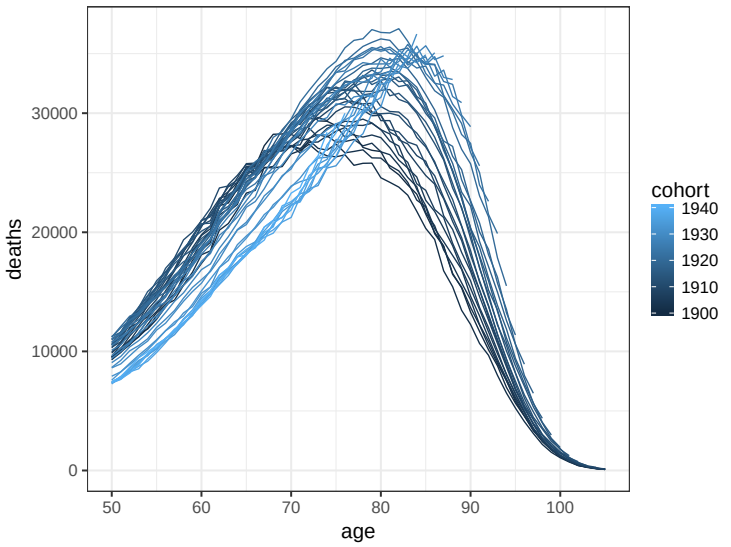


Figure 4.10: Death counts by age, United States, males, cohorts 1900-1940, ages 50-105. Each line is a different cohort. Data come from the HMD.

able in the HMD and the resulting fits based on both the Gompertz and principal component approaches. The performance of the two methods is compared based on several in- and out-of-sample diagnostic measures.

4.7.1 Data

The two models are fit to HMD data for U.S. males for cohorts 1900-1940, for ages 50-105. In order to fit to comparable data available in CenSoc, the cohort-based death rates and exposure to risk by age are converted into implied death counts by age. Fig. 4.10 shows death counts by age and cohort. While data on the deaths distribution is complete for older, already extinct cohorts, only part of the deaths distribution is available for the younger cohorts.

4.7.2 Computation

The hierarchical model frameworks specified above were fit within Bayesian frameworks using the statistical software R. Samples were taken from the posterior distributions of the parameters via a Markov Chain Monte Carlo (MCMC) algorithm. This was performed using JAGS software (Plummer, 2003). Standard diagnostic checks using trace plots and the Gelman and Rubin diagnostic (Gelman and Rubin, 1992) were used to check convergence.

For the principal components approach, initial values $\beta^*_{d,c}$ for the coefficients on the principal components were chosen to ensure that $\sum_x d^*_{c,x} = 1$. These were obtained

by first running the model without constraints to get an idea of plausible coefficient estimates. Initial values were chosen such that

$$\text{logit}^{-1}\left(\sum_x P_{0,x} + \beta^*_{1,c}P_{1,x} + \beta^*_{2,c}P_{2,x}\right) = 1$$

Best estimates of all parameters of interest were taken to be the median of the relevant posterior samples. The 95% Bayesian credible intervals were calculated by finding the 2.5% and 97.5% quantiles of the posterior samples.

4.7.3 Converting estimates to other measures of mortality

In both the Gompertz and principal components approaches, we obtain samples from the estimated posterior distribution of d^*_x, i.e. the (truncated) deaths distribution across age. These quantities can be converted into other mortality indicators, such as life expectancy at age 50, by utilizing standard relationships between life table quantities (Preston et al. (2000); Wachter (2014)). In particular, the proportion of people surviving to each age, l_x, is calculated using reverse survival,

$$l_x = 1 - \sum_{x+1}^{\omega} d^*_x$$

i.e. the proportion alive at age x is 1 minus the sum of those who died in age groups above age x, where ω is the last age group (in this case, 105). The person-years lived between ages x and $x+1$, i.e. L_x is estimated as

$$L_x = \frac{l_x + l_{x+1}}{2}.$$

The person-years lived above age x is then

$$T_x = \sum_x L_x$$

and the life expectancy at age x is then

$$e_x = \frac{T_x}{l_x}.$$

The above life table relationships are calculated based on all samples from the posterior distribution of d^*_x, resulting in a set of samples for e_x. The corresponding 95% credible intervals around the estimates of e_x can be calculated based on the 2.5th and 97.5th percentiles of the samples.

4.7.4 Gompertz results

Results from fitting the truncated Gompertz hierarchical model are shown in Figs. 4.11-4.13. The mode age of death is steadily increasing over time, from around age 76

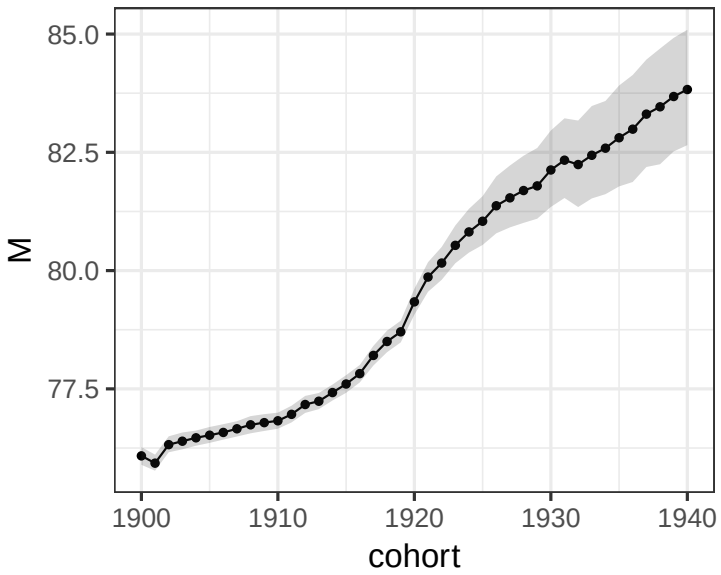


Figure 4.11: Estimates of Gompertz mode age of death, United States. Median posterior estimates are shown by the dots. The shaded area represents the 95% uncertainty interval.

in cohort 1900 to 84 in cohort 1940 (Fig. 4.11). In terms of the Gompertz slope parameter, after remaining fairly constant in the earlier cohorts, the estimate for β decreases across cohorts 1915-1925. Since 1925, however, the estimated values of β have stagnated.

The uncertainty intervals around the estimates for both M and β increased for the younger cohorts. This reflects the fact that less data are available in the cohorts. For example, for the 1940 cohort, observed death counts in HMD are only available up to age 74. As such, the model is fitting a deaths distribution across all ages based on only partial information about the shape of the distribution from the data.

Fig. 4.13 illustrates the fitted death distributions in comparison with the available data for nine cohorts between 1900-1940. In general, the truncated Gompertz model captures the main characteristics of the shape of the distributions well, as well as changes across cohorts. In the older cohorts in particular, the Gompertz curve is not an exact fit to the HMD data, and seems to place too much mass around the mode age of death, and not enough mass on younger ages of death (60-70). For cohorts younger than the 1925 cohort, the model is fitting the full curve based on only having data on the left side, with no real information about the modal age of death. However, fits for these cohorts are partially informed by past cohorts, through the temporally-correlated prior that was placed on M.

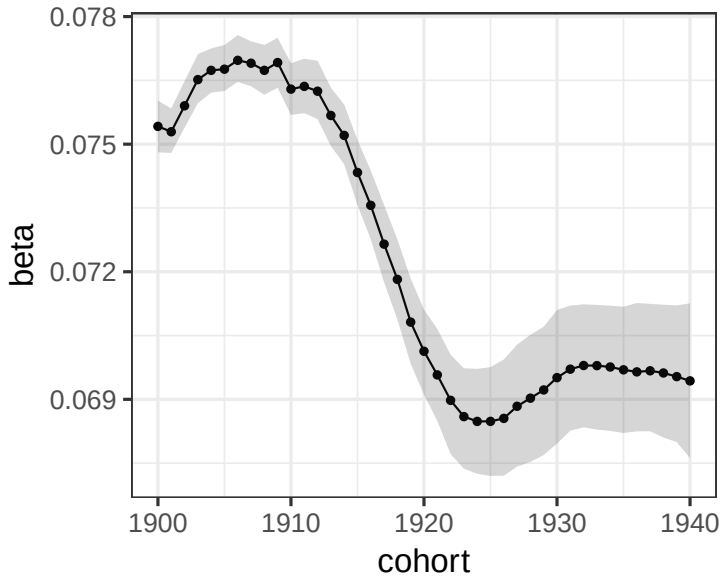


Figure 4.12: Estimates of Gompertz β, United States. Median posterior estimates are shown by the dots. The shaded area represents the 95% uncertainty interval.

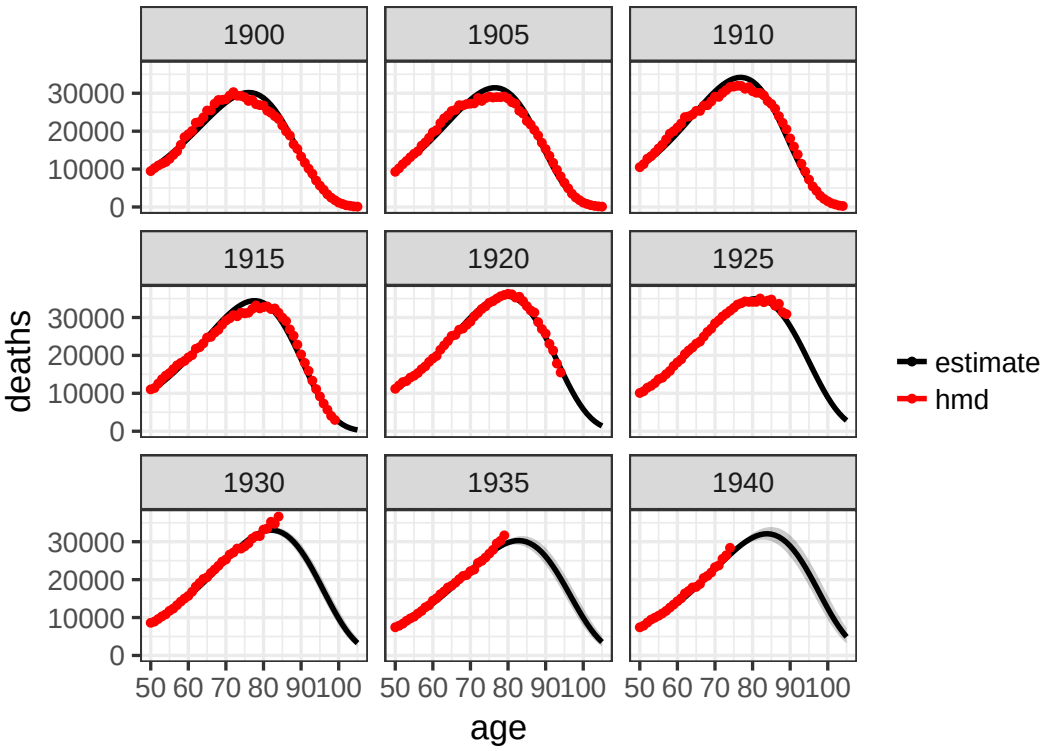


Figure 4.13: Truncated Gompertz model estimates and HMD data of deaths by age for nine cohorts between 1900 and 1940. Data from HMD are shown by the red dots. The estimates and associated 95% uncertainty intervals are shown by the black lines and shaded areas.

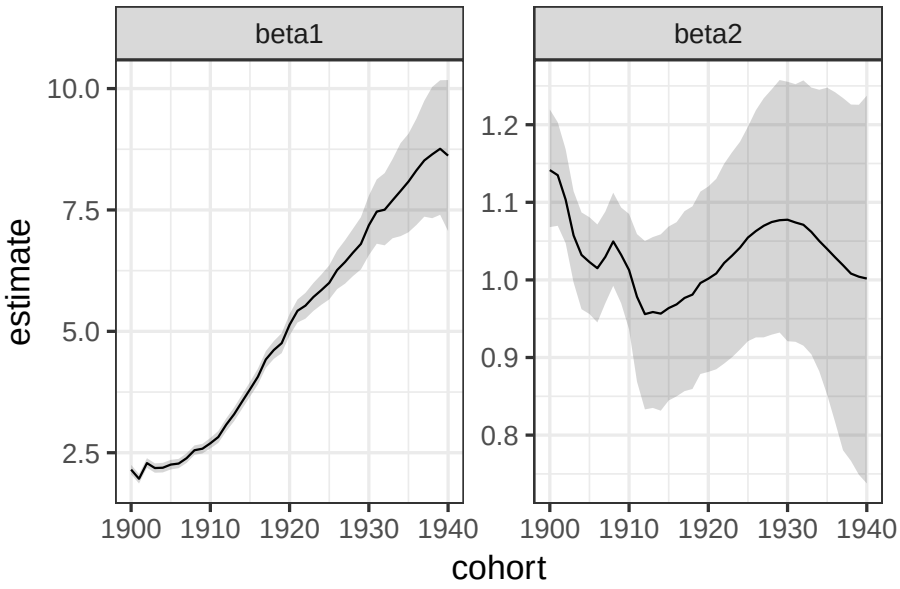


Figure 4.14: Estimates of principal component coefficients $\beta_{1,c}$ (left) and $\beta_{2,c}$ (right) across cohorts. Median posterior estimates are shown by the black lines. The shaded areas represent the 95% uncertainty intervals.

4.7.5 Principal component regression results

Results from fitting the principal components regression model are shown in Fig. s 4.14 and 4.15. The coefficient on the first principal component steadily increased over cohorts (Fig. 4.14). This represents a shift in the mass of the deaths distribution away from younger ages and towards the older ages. The coefficient on the second principal component broadly decreased over cohorts, but remained positive. Note that the uncertainty around the coefficient estimates increases across cohorts, as less information about the shape of the deaths distribution is available.

Fig. 4.15 illustrates the fitted death distributions in comparison with the available data for nine different cohorts between 1900-1940. In general, the principal components model seems to produce fairly similar fits to the Gompertz model. However, especially in younger cohorts, the uncertainty around estimates is larger.

4.7.6 Comparison of models

Figs. 4.16 illustrates the estimates of the hazard rate at each age x on the log scale for the two models. A key assumption of the Gompertz model is that hazards are assumed to be log-linear, which is illustrated by the estimates in the left-hand panel. In contrast, the estimated hazards from the principal components model are not quite log-linear, with evidence of an increasing slope at older ages. For both models, hazards are decreasing across cohort.

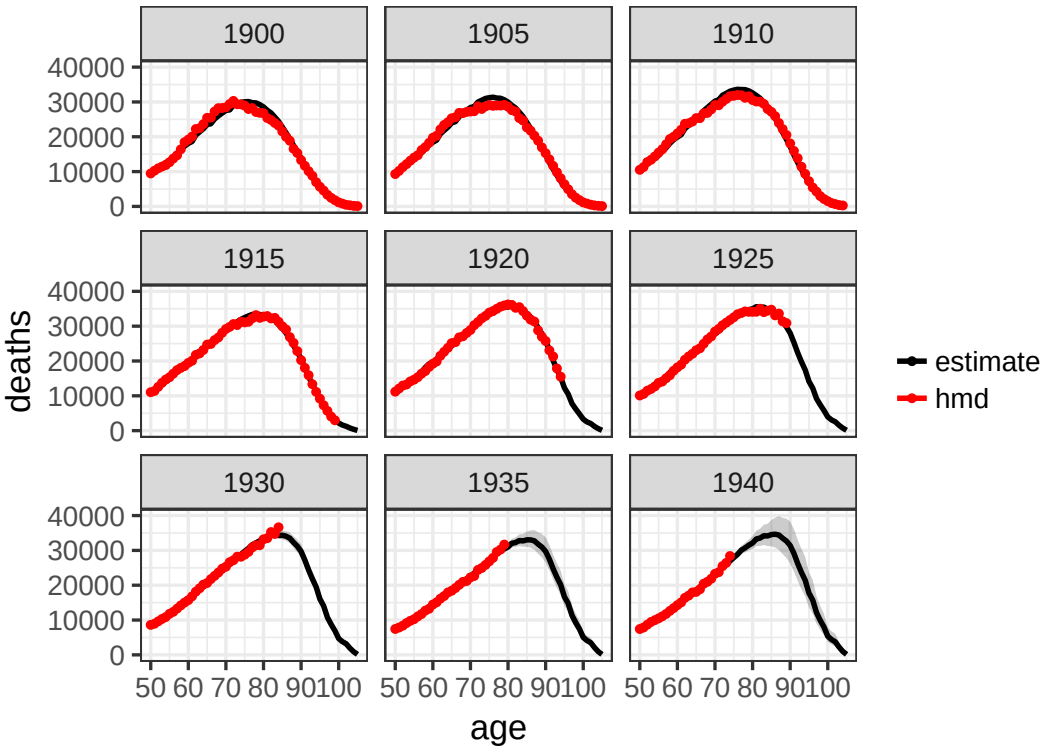


Figure 4.15: Principal component estimates and HMD data of deaths by age for nine cohorts between 1900 and 1940. Data from HMD are shown by the red dots. The estimates and associated 95% uncertainty intervals are shown by the black lines and shaded areas.

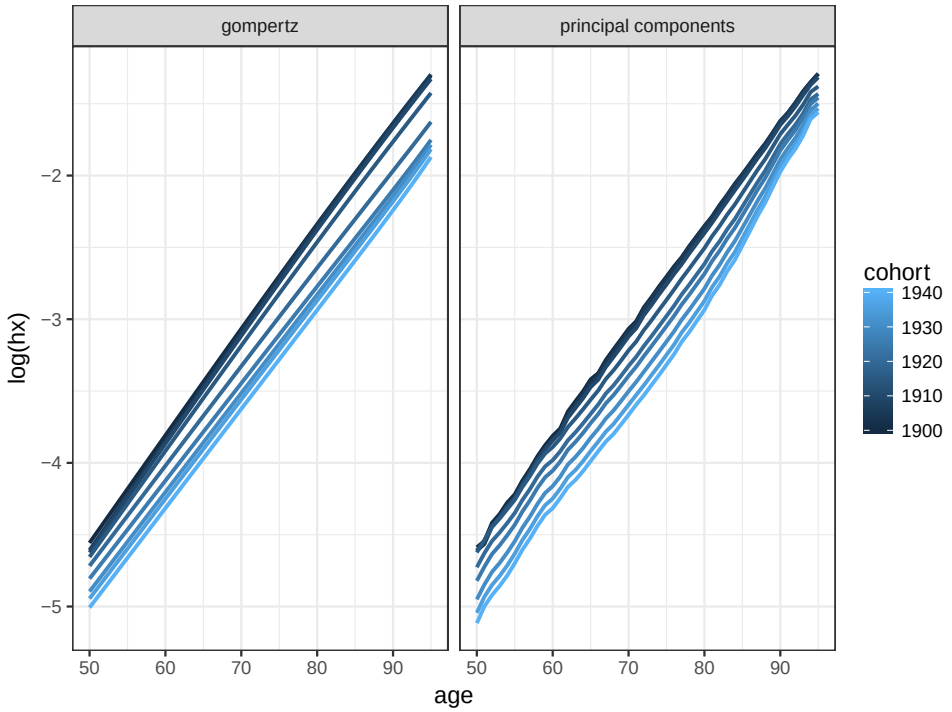


Figure 4.16: Estimated (log) hazard rates by cohort for the truncated Gompertz (left) and principal components model (right). Each line represents a different cohort.

Fig. 4.17 shows the estimates of life expectancy at age 50 (e_{50}) across cohorts for the two models. The estimates are quite similar across the two models for earlier cohorts, but start to diverge around the 1920 cohort, where there is lessening information available about the shape of the mortality curve. However, the estimates start to converge again in more recent cohorts, and there is no significant difference between the estimates by 1940. The uncertainty around the principal components is slightly larger in later cohorts.

Model Performance

Several measures are considered to compare the performance of the Gompertz and principal components models based on estimates of U.S. mortality using HMD.

Firstly, the relative performance of the models was assessed using the Watanabe-Akaike or widely available information criterion (WAIC), which measures a combination of model fit and a penalty based on the number of parameters (Vehtari et al. (2017)). The lower the WAIC, the better the model. The Gompertz model resulted in a WAIC of -3876 compared to -4309 for the principal components model. Thus, based on this measure, the principal components model outperforms the Gompertz model.

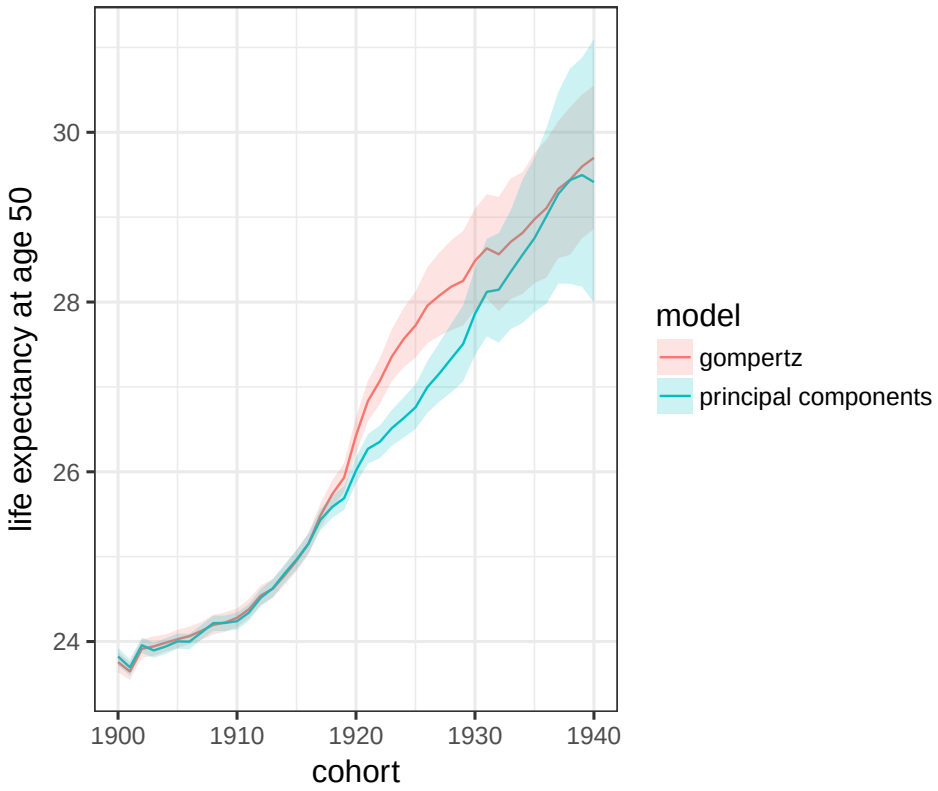


Figure 4.17: Estimated life expectancy at age 50 by cohort for the Gompertz (red line) and principal components (blue line) models. The median estimates are shown as the lines, and 95% uncertainty intervals are shown by the shaded areas.

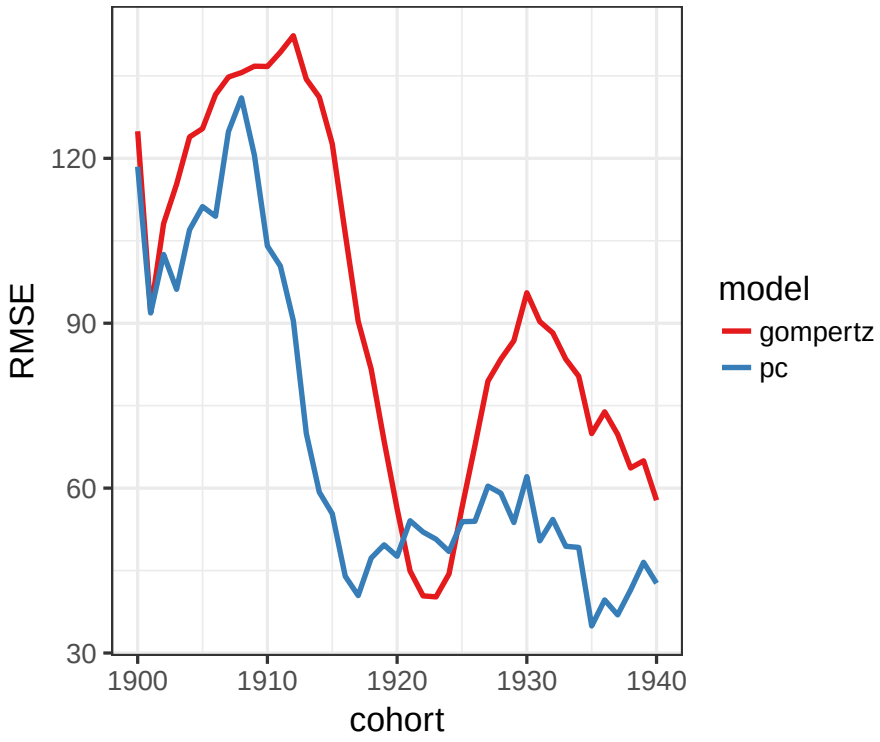


Figure 4.18: RMSE by cohort for the Gompertz (red line) and principal components (blue line) models.

Secondly, the root mean squared error (RMSE) of fitted values compared to HMD values was estimated, across both age and cohort. RMSE across cohorts is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{A}\sum_{x=1}^{A}(\hat{y}_{c,x} - y^*_{c,x})^2}, \tag{4.24}$$

where $\hat{y}_{c,x}$ is the estimated death count at age x for cohort c, $y^*_{c,x}$ is the true mortality rate and A is the number of ages. In a similar way, the RMSE across age is

$$\text{RMSE} = \sqrt{\frac{1}{C}\sum_{c=1}^{C}(\hat{y}_{c,x} - y^*_{c,x})^2}, \tag{4.25}$$

where C is the number of cohorts. Figs. 4.18 and 4.19 plot the RMSE across cohort and age for each model. In terms of both cohort and age, for the most part, the principal components model has a lower RMSE.

One final measure that was considered to compare the two models was the coverage of the prediction intervals. Given that observed death counts by age are distributed

$$y_{c,x} \sim Poisson(\lambda_{c,x})$$

new observations of deaths by age and cohort, $y^{new}_{c,x}$ can be predicted based on this distribution. Repeating this simulation many times gives a posterior predictive distribution of $y_{c,x}$. Prediction intervals can be calculated based on this distribution

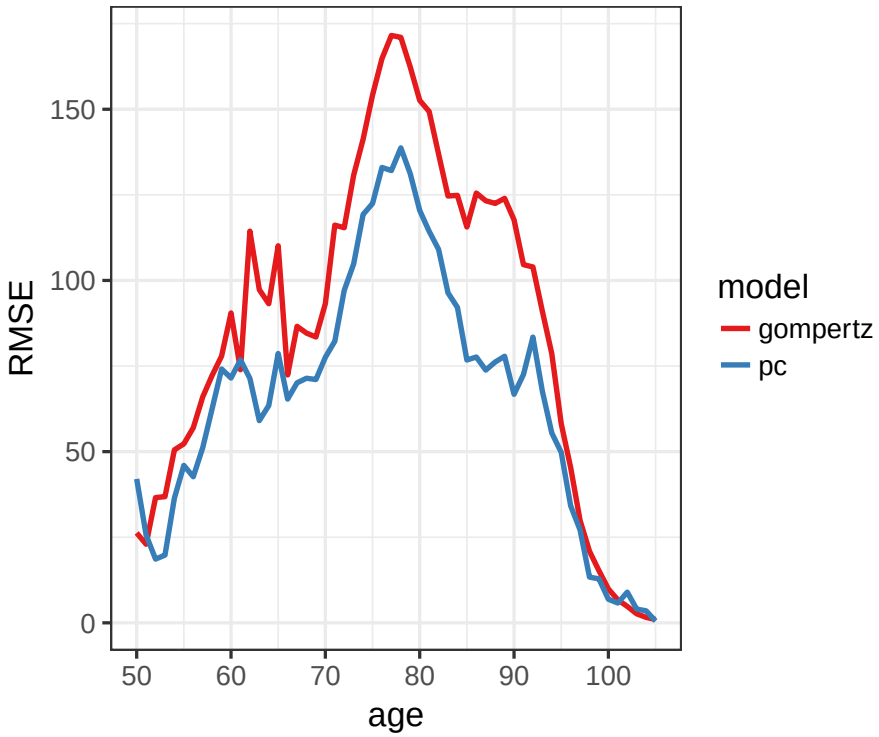


Figure 4.19: RMSE by age for the Gompertz (red line) and principal components (blue line) models.

and the coverage of such intervals assessed. For example, we would expect 95% prediction intervals of $y_{c,x}^{new}$ to include the observed values $y_{c,x}$ 95% of the time.

Fig. 4.20 illustrates the coverage of 95% prediction intervals across cohorts for both models. Coverage of the intervals for both models are lower than expected for the earlier cohorts; however, from around the 1915 cohort, the coverage is at least 95%. This suggests the uncertainty intervals are reasonably well calibrated.

4.7.7 Discussion

Both models fit reasonably well to HMD data, capture the main patterns in the death distribution and how it changes across cohorts. These models illustrate how underlying demographic structures can be fit within a Bayesian framework to get plausible estimates of death distributions when only truncated data are available. The principal components model appears to slightly out-perform the Gompertz model across several different measures. In particular, the WAIC and RMSE measures were lower for the principal components model, suggesting that it does a better job at fitting to the HMD data.

The advantage of the principal components approach is that the underlying mortality structure is determined from real mortality data across a wide range of populations

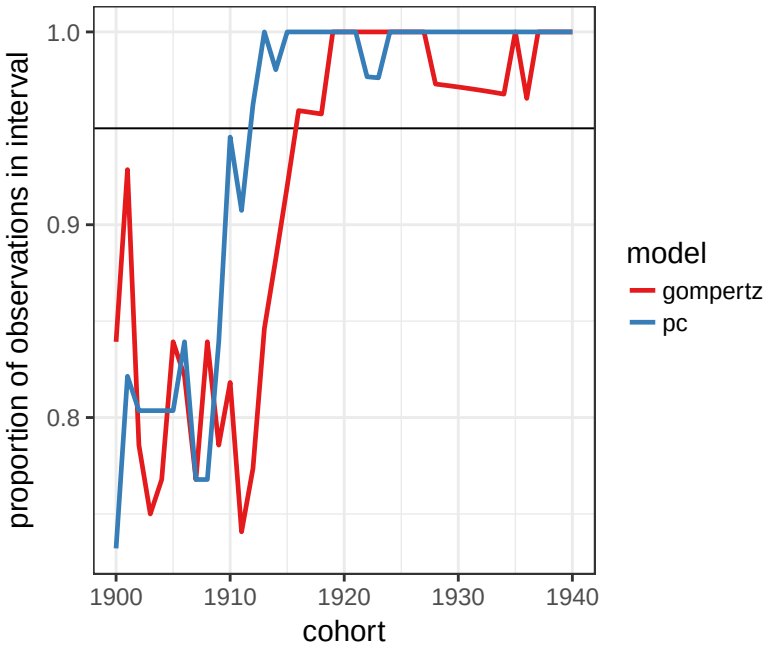


Figure 4.20: Coverage of 95% prediction intervals for the Gompertz (red line) and principal components (blue line) models. If the uncertainty intervals are well-calibrated, the coverage of the prediction intervals would be expected to be 95%.

and time periods. The model is more flexible and better able to fit to death distributions that do not follow a simple parametric form. Thus, complex patterns in mortality data can be captured with relatively few parameter inputs. However, from a computational perspective, the principal components model requires initial conditions to be chosen to satisfy the constraint on the death distribution. In addition, fitting the principal components method requires extra data processing to obtain usable principal components, and decisions need to be made about the appropriate mortality standard. While non-U.S. HMD data was used as standard, potentially any standard could be chosen, and more than two principal components could be included into the model.

While the Gompertz model did not statistically perform quite as well as the principal components model, it still has the advantage of being a well-known, simple parametric model. Gompertz parameters are easily interpreted and can be compared across different populations and studies. There is no requirement for a particular mortality standard to be chosen and justified. In summary, there are advantages and disadvantages to both methods, and model performance is reasonable for both options.

4.8 Estimating mortality inequalities using CenSoc

In this section, the mortality modeling approaches discussed above are applied to the CenSoc dataset to estimate mortality outcomes across cohorts and socioeconomic status (SES). In particular, mortality differences are estimated across education and income groups.

Given the relative performance of the two modeling approaches in fitting to the HMD data, the principal components model is used to estimate the death distributions by cohort and SES. This approach appears to offer slightly more flexibility in fitting to, and capturing, the main characteristics of the partially observed death distributions.[4] The method can be extended to allow for differing mortality trends by socioeconomic group, as shown below.

4.8.1 Mortality trends by education group

Education can affect mortality outcomes through a variety of different pathways (Hummer et al. (1998); Elo (2009)). Education may indirectly affect mortality and health outcomes through being associated with higher income, thereby increasing an individual's available resources to spend on health. Greater access to education also allows individuals to make more informed decisions about their health and lifestyle choices. Education may also mean increased social support, less exposure to acute and chronic stress, and a greater cognitive ability to cope with stressful situations.

As an SES measure, education has the advantage of having temporally stable defined categories over time. In addition, unlike income or occupation, education changes very little over the lifecourse. It reflects the stock of human capital established relatively early in life that is available to individuals throughout their life course (Elo (2009)).

We use CenSoc to estimate the relationship of years of schooling and mortality across cohorts. The 1940 census contains information on the number of years of schooling, from zero to 17+ years. The number of years of schooling was recoded into six levels:

- less than middle school (less than 8 years)
- middle school (8 years)
- some high school (8-11 years)
- high school (12 years)

[4]Note that the Gompertz approach was also fitted to the CenSoc data by SES group, with the resulting estimates being very similar to those produced by the principal components method.

- some college (13-15 years)
- college or more (16+)

The analysis includes the 25 birth cohorts 1890-1915, meaning the respondents were at least 25 years old at the time of the census.

The principal components modeling framework described in Section 6 is extended to allow the principal component coefficients β_1 and β_2 to vary not only by cohort c but also by education level g. As before, the mean death distribution and the two principal components were derived from cohort-based HMD data across all available countries. The principal component coefficients β were modeled using random walks across cohorts for each education group. The full model is:

$$\begin{aligned}
D_{c,g} &\sim Poisson(N_{c,g}) \\
y_{c,g,x} &\sim Poisson(\lambda_{c,g,x}) \\
\lambda_{c,g,x} &= N_{c,g} \cdot d^*_{c,g,x} \\
\text{logit } d^*_{c,g,x} &= P_{0,x} + \beta_{1,c,g} P_{1,x} + \beta_{2,c,g} P_{2,x} \\
d^*_{c,g} &= \sum_x d^*_{c,g,x} = 1 \\
\beta_{d,c,g} &\sim N(2\beta_{d,g,c-1} - \beta_{d,g,c-2}, \sigma^2_{d,g}) \text{ for } d = 1 \\
\beta_{d,c,g} &\sim N(\beta_{d,g,c-1}, \sigma^2_{d,g}) \text{ for } d = 2 \\
\sigma_{d,g} &\sim U(0, 40)
\end{aligned}$$

Estimates and uncertainty for life expectancy at age 50 can be obtained using samples from the estimated posterior distribution of $d^*_{c,g,x}$. Life expectancy at age 50 is calculated for each cohort and education group, i.e. $e_{50,c,g}$.

Results

Fig. 4.21 shows the distribution of deaths by age and education level for different cohorts. The available data is shown by the dots, and the resulting estimate and 95% credible intervals are shown by the colored lines and associated shaded area. This figure illustrates the changing distribution of education across cohorts. In the older cohorts, the largest groups were those who had less than a high school education. Over time the largest group becomes those with a high school certificate. Fig. 4.21 also illustrates the differing amounts about ages of death information available by cohort. For the older cohorts, we observe the deaths at older ages, while the opposite is true for younger cohorts. Thus, moving through cohorts we observe the shape of the death distribution on the right, moving to the left.

Estimates and 95% uncertainty intervals for life expectancy at age 50 by education group are shown in Fig. 4.22. In general, mortality disparities between the least and

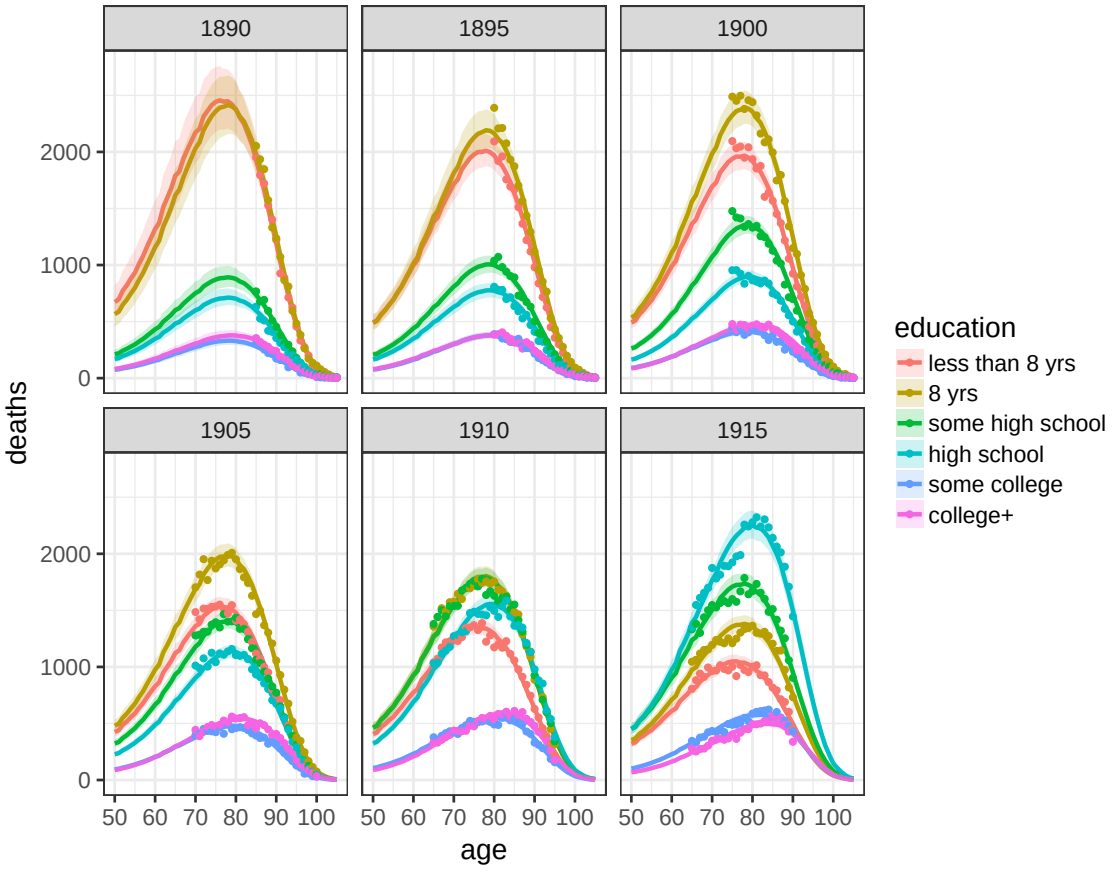


Figure 4.21: Estimated and observed death counts in CenSoc by education level, cohorts 1890-1915. The available data are shown by the dots, the estimates and associated 95% uncertainty intervals are shown by the lines and shaded areas.

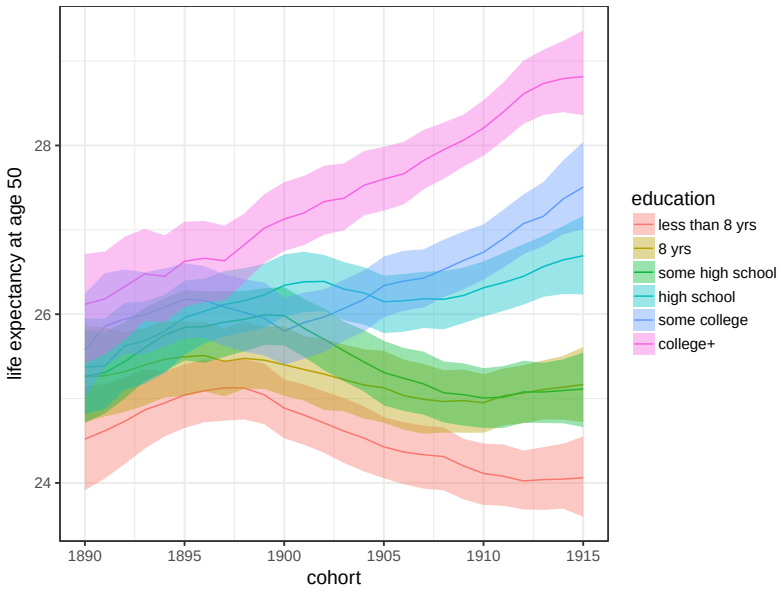


Figure 4.22: Life expectancy at age 50 by education level, cohorts 1890-1915. The shaded areas represent 95% uncertainty intervals.

most-educated groups are increasing over time. For the older cohorts, life expectancy appeared to be generally increasing for all education groups, with no significant difference in the estimates; however, since around 1900 there has been a divergence in outcomes.

For those in the education groups who had a high school certificate or higher, life expectancy increased across cohorts. For example, e_{50} those who had a college degree or higher increased from around 26.2 years to 28.5 years over the cohorts 1890-1915. The e_{50} is consistently around 1 year lower for those who had at least some post high school education but had not completed the college degree. The e_{50} for the high school group also increased over cohorts, although there was a period of stagnation between cohorts 1900-1915. There is no significant difference in the life expectancy for those with high school only and those with some post high school education.

For those population groups with less than high school, life expectancy stagnated or declined. Interestingly, there is very little difference in e_{50} for those who have 8 years of schooling compared to those who have some high school education. For the 1915 cohort, the estimate of e_{50} for these groups was around 3.5 years less than the most educated group. Life expectancy for those with less than middle school education initially increased, but has declined over time since around cohort 1897.

These results are broadly consistent with previous research which observes increasing disparities in mortality across education over time. Previous research has illustrated the clear education gradient in mortality that exists in the United States, with most studies finding the mortality differential between the lowest and highest education groups has increased over time (Masters et al. (2012); Hummer and Hernandez (2013); Hendi (2015); Krueger et al. (2015)). Fig. 4.22 suggests the widening dispar-

ity is a consequence of both increases in the higher education groups, and decreases in the lower-educated groups. As illustrated in Fig. 4.21, the least-educated groups are decreasing in size over cohort, and those left in the lowest education group may be becoming a more selective group with relatively worse outcomes.

4.8.2 Mortality trends by income

Mortality outcomes are strongly associated with income level, both at the individual and aggregate levels (Preston (1975); Deaton and Paxson (2001)). Higher income can improve mortality in the absolute sense, through the availability of greater resources. Individual income may be important in the relative sense, through its relation to determining social class. For example, Wilkinson and Pickett argue that income mostly affects health and mortality through the psychosocial factors associated with an individual's relative position in the social hierarchy (Wilkinson (2006); Pickett and Wilkinson (2015)).

The main measure of income in the 1940 census is the respondent's total pre-tax wage and salary income for the previous year. Thus, there is only one observation of each person's income in 1940, and so the mortality analysis by income group is based on this historical measure of income, rather than income near the time of death. As the observation window of deaths is 1975-2005, the income measure is at least 35 years prior to death. This has the disadvantage of being outdated information and not reflective of an individual's wealth at a time closer to death. However, historical income is mostly likely strongly correlated with more recent income, and so this measure is still a proxy for more recent SES. In addition, the historical measure of income may be less subject to reverse causality issues, that is, the possibility that someone who is suffering from a serious illness is unlikely to be working full time.

The analysis above is repeated by broad income group. Income groups are defined based on the quartile of an individual's income from wages in 1940 for the relevant age group. Those with zero income are removed from the analysis, as they are likely to be self-employed and include many high income individuals. The same modeling framework defined in the previous section was fit to birth cohorts 1890-1915, divided into the four income groups. Life expectancy at age 50 for each cohort and income group, i.e. $e_{50,c,g}$, was again calculated based on the posterior samples of the truncated death distribution $d^*_{c,g,x}$.

Results

Fig. 4.23 shows the estimates and 95% uncertainty intervals for life expectancy at age 50 by income group across the 25 birth cohorts. Up until around 1900, there was no significant difference in the estimates across income groups. For cohorts younger than 1900, however, life expectancy has increased for the two highest groups, i.e. those

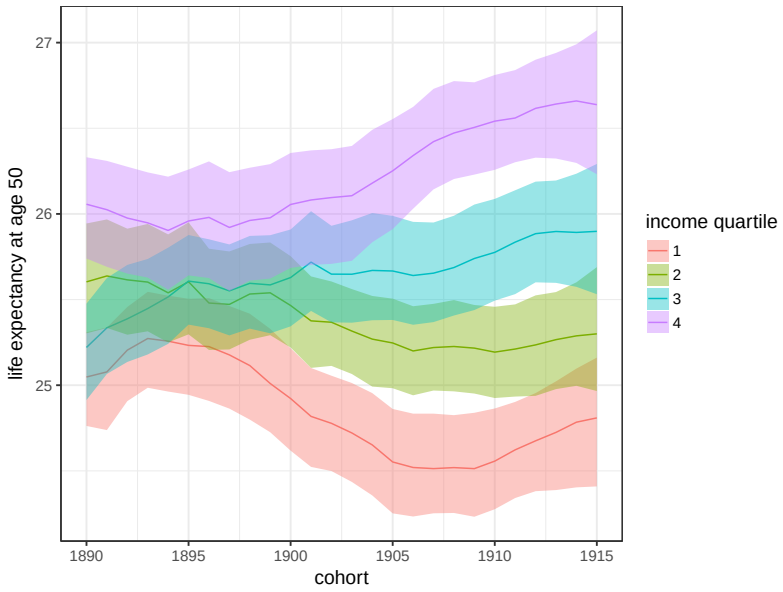


Figure 4.23: Life expectancy at age 50 by income group, cohorts 1890-1915. The shaded areas represent 95% uncertainty intervals.

who had higher than median income, and decreased for those groups below median income, such that the life expectancy gap in the 1915 cohort was around 1.5 years. While the higher income groups experienced improving mortality conditions, life expectancy in the lower income showed some evidence of stagnation or decline. The lowest income group had declining life expectancy until cohort 1905, although this has stagnated in more recent cohorts. In general, mortality inequality has increased over time. This observation is broadly consistent with other studies, which find evidence for increasing mortality inequality across income at both the individual and county levels, and a stagnation of progress in the lower income groups (Waldron (2007); Bosworth and Burke (2014); Chetty et al. (2016); Currie and Schwandt (2016)).

4.9 Discussion

This paper described 'CenSoc', a dataset which was created using the 1940 U.S. Census and Social Security Deaths Masterfile, and contains over 7.5 million records that link individual's demographic, socioeconomic and geographic information with their age and date of death. This dataset is available for researchers to study mortality disparities and changes in the United States.

In contrast to many studies of socioeconomic inequalities in mortality, which use data at the aggregate level, CenSoc provides information at the individual level, so that mortality outcomes can be directly related to SES. The large number of records available in CenSoc provides greater statistical power in analysis compared

to other smaller linked datasets such as the NHIS Linked Mortality Files or the National Longitudinal Mortality Study. In addition, the mortality estimates do not rely on comparing deaths data from one source to population data from another source, thereby avoiding common problems associated with numerator-denominator bias when studying mortality by race, for example (Preston and Elo (1999); Black et al. (2017)).

By construction, CenSoc only contains individuals that 1) died in the period 1975-2005; and 2) were successfully matched across the two datasets. As a consequence, the mortality information available in CenSoc is of the form of left- and right-truncated deaths by age, with no information about the relevant population at risk at any age or cohort. This means that, apart from extinct cohorts (those that have reached an age in 2005 where there are very few or no survivors), standard techniques of survival analysis and mortality estimation cannot be used. As such, a second part of this paper was to develop mortality estimation methods in order to best to utilize the 'deaths without denominators' information contained in CenSoc. Two methods of estimating the truncated deaths distribution across age, cohort (and potentially other population subgroups) were presented. Both methods were fit in a Bayesian setting, which allowed for the incorporation of priors on parameters of interest, and also allowed for uncertainty in the resulting estimates and other quantities of interest, such as life expectancy, to be reported, with minimal additional calculations.

The first was a parametric approach, modeling death distributions under the assumption of Gompertz hazards. A particular contribution of this approach was to formulate the Gompertz model as a Bayesian hierarchical framework, which allowed somewhat informative priors to be placed on the mode age at death and how it changed over time. Placing structure on trends over cohort accounts for autocorrelation in mortality trends and has the additional benefit of providing a clear framework for the projection of mortality trends into the future. Results of fitting the truncated Gompertz model to HMD data for the United States gave reasonable results. However, discrepancies between the fitted and observed death distributions, as well as the relatively narrow uncertainty intervals around the estimates, highlighted the inherent lack of flexibility of the Gompertz approach.

The second modeling approach presented was the principal components model. In contrast to a purely parametric model, this approach extracts key patterns from high-quality mortality data and uses them as the basis of a regression for the death distributions over age and cohort. The death distributions are thus modeled as a combination of these key mortality patterns, and changes in the relative combinations over time are estimated within a Bayesian hierarchical framework. The principal components approach offered a greater flexibility compared to the Gompertz model, and generally produced lower root mean squared errors and better coverage of uncertainty intervals when fit to HMD data.

After testing on HMD data, the principal components model was then used to estimate mortality outcomes by education and income using the CenSoc dataset. Dif-

ferences by group were estimated across 25 birth cohorts from 1890 to 1975. Results suggest that both the education and income gradient in adult mortality has increased over time. This is a consequence not only of life expectancy in the highest education/income groups increasing at a faster rate, but also of life expectancy in the lowest SES groups stagnating, or even declining in the case of education.

There are several limitations of this work and areas for future research. Firstly, while the resulting CenSoc dataset is quite large in terms of absolute numbers, the raw match rate is only around 20%, (and at its highest, around 28% for 20-24 year olds). While some individuals are unmatched for mortality reasons — i.e. dying outside of the SSDM window — others are missed because of the matching method. Indeed, the method used to match across the two datasets is very simple, using only exact matches of name and age, and only taking unique combinations. Any duplicate keys are discarded. In addition, the exact match process does not account for small changes in name across the two datasets, such as spelling errors, the use of nicknames, or the use of initials. Future work will focus on more detailed data cleaning to pick up name errors, and also investigating probabilistic matching techniques to deal with partial matches, based on Jaro-Winkler distances (Jaro (1989); Winkler (1990)) or NYIIS phonetic codes (Abramitzky et al. (2018)). Another option would be to create multiple datasets based on duplicate keys and calculate mortality indicators across all datasets, with the related uncertainty in estimates constructed using bootstrapping techniques.

The CenSoc dataset contains only males. Matching females across the two datasets is more difficult because of the possibility of marriage between the time of the census and SSDM window, which would lead to a name change. As the SSDM only contains information on name, date of birth and date of death, there is no information about marital status at time of death. It would be possible to create a dataset with married females (married at the time of census), and to investigate partial matches based on age, first and middle name, and probable marriage rates.

For the principal components approach, the age distribution was cut off at a maximum age of 105. This was chosen mostly because of observed discontinuities in the principal components derived from HMD data, with large, sudden changes in the principal components at older ages. Discontinuities may partly be an artifact of the methods of estimate used by HMD, which switch over at around age 80-90, depending on the country and quality of raw data available (Wilmoth et al. (2007)). Future work will investigate the sensitivity of principal components to choices of different countries and cohorts being included in the matrix on which the SVD is performed.

Notwithstanding these areas for future research, the CenSoc project provides a useful new data source for the study of mortality disparities and change over time. This paper introduced the dataset and also developed methods to fully utilize the mortality information available. The CenSoc data, code used to match the raw data, code and functions to estimate mortality indicators, and the relevant documentation has been made publicly available (at the time of writing at: `https:`

//censoc.demog.berkeley.edu/). By providing an open source, transparent resource, the goal is to encourage reproducibility of research and provide a resource for other researchers to help answer their own research questions of interest.

References

Abramitzky, R., R. Mill, and S. Pérez (2018). Linking Individuals Across Historical Sources: a Fully Automated Approach. Technical report, National Bureau of Economic Research.

Alexander, M. and L. Alkema (2018). Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model. *Demographic Research 38*, 335–372.

Alexander, M., E. Zagheni, and M. Barbieri (2017). A flexible Bayesian model for estimating subnational mortality. *Demography 54*(6), 2025–2041.

Alexander, R. and Z. Ward (2018). Age at Arrival and Assimilation during the Age of Mass Migration. *Journal of Economic History*, Forthcoming.

Alho, J. and B. Spencer (2006). *Statistical demography and forecasting*. Springer Science & Business Media.

Alkema, L., D. Chou, D. Hogan, S. Zhang, A.-B. Moller, A. Gemmill, D. M. Fat, T. Boerma, M. Temmerman, C. Mathers, and L. Say (2016). Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group. *The Lancet 387*(10017), 462–474.

Alkema, L., V. Kantorova, C. Menozzi, and A. Biddlecom (2013). National, regional, and global rates and trends in contraceptive prevalence and unmet need for family planning between 1990 and 2015: a systematic and comprehensive analysis. *The Lancet 381*(9878), 1642–1652.

Alkema, L. and J. R. New (2014). Global estimation of child mortality using a Bayesian B-spline bias-reduction model. *The Annals of Applied Statistics 8*(4), 2122–2149.

Alkema, L., A. E. Raftery, P. Gerland, S. J. Clark, F. Pelletier, T. Buettner, and G. K. Heilig (2011). Probabilistic projections of the total fertility rate for all countries. *Demography 48*(3), 815–839.

Alkema, L., M. B. Wong, and P. R. Seah (2012). Monitoring Progress Towards Millennium Development Goal 4: A Call for Improved Validation of Under-Five Mortality Rate Estimates. *Statistics, Politics and Policy 3*(2), 1–19.

Andreev, K., D. Jdanov, E. Soroko, and V. Shkolnikov (2003). Methodology: Kannisto Thatcher database on old age mortality. *Max Planck Institute for Demographic Research, Rostock, Germany.[Online]*.

Azose, J. J., H. Ševčíková, and A. E. Raftery (2016). Probabilistic population projections with migration uncertainty. *Proceedings of the National Academy of Sciences 113*(23), 6460–6465.

Barbieri, M. and R. Depledge (2013). Mortality in France by département. *Population 68*(3), 375–417.

Bergeron-Boucher, M.-P., M. Ebeling, and V. Canudas-Romo (2015). Decomposing changes in life expectancy: Compression versus shifting mortality. *Demographic Research 33*, 391–424.

Bhutta, Z. A., M. Chopra, H. Axelson, P. Berman, T. Boerma, J. Bryce, F. Bustreo, E. Cavagnero, G. Cometto, B. Daelmans, A. de Francisco, H. Fogstad, N. Gupta, L. Laski, J. Lawn, B. Maliqi, E. Mason, C. Pitt, J. Requejo, A. Starrs, C. G. Victora, and T. Wardlaw (2010, 2017/07/27). Countdown to 2015 decade report (2000–10): Taking stock of maternal, newborn, and child survival. *The Lancet 375*(9730), 2032–2044.

Bijak, J. (2008). Bayesian methods in international migration forecasting. *International migration in Europe: Data, models and estimates*, 255–288.

Bijak, J. (2010). *Forecasting international migration in Europe: A Bayesian view*, Volume 24. Springer Science & Business Media.

Bijak, J. and J. Bryant (2016). Bayesian demography 250 years after Bayes. *Population Studies 70*(1), 1–19. PMID: 26902889.

Black, D. A., Y.-C. Hsu, S. G. Sanders, L. S. Schofield, and L. J. Taylor (2017). The Methuselah Effect: The Pernicious Impact of Unreported Deaths on Old-Age Mortality Estimates. *Demography 54*(6), 2001–2024.

Booth, H., R. J. Hyndman, L. Tickle, P. De Jong, et al. (2006). Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions.

Bosworth, B. and K. Burke (2014). Differential mortality and retirement benefits in the Health and Retirement Study. *Center for Retirement Research Working Paper 2014*(4).

Bradley, S. (2015). More Questions, More Bias? An Assessment of the Quality of Data Used for Direct Estimation of Infant and Child Mortality in the Demographic and Health Surveys. Paper presented at PAA 2015. Available at `http://paa2015.princeton.edu/uploads/152375`.

Brass, W. (1971). On the scale of mortality. In *Biological aspects of demography*, pp. 69–110. Taylor & Francis.

Bravo, J. and J. Malta (2010). Estimating life expectancy in small population areas. Joint Eurostat/UNECE Work Session on Demographic Projections Working Paper 10.

Burger, O. and T. I. Missov (2016). Evolutionary theory of ageing and the problem of correlated Gompertz parameters. *Journal of Theoretical Biology 408*, 34–41.

Camarda, C. G. (2012). MortalitySmooth: An R package for smoothing Poisson counts with P-splines. *Journal of Statistical Software 50*(1), 1–24.

Canudas-Romo, V. (2008). The modal age at death and the shifting mortality hypothesis. *Demographic Research 19*, 1179–1204.

Case, A. and A. Deaton (2017). Mortality and morbidity in the 21st century. *Brookings papers on Economic Activity 2017*, 397.

Centre for Research on the Epidemiology of Disasters (CRED) (2012). EM-DAT: The CRED International Disaster Database. Available at: `http://www.emdat.be/database`.

Chetty, R., M. Stepner, S. Abraham, S. Lin, B. Scuderi, N. Turner, A. Bergeron, and D. Cutler (2016). The association between income and life expectancy in the United States, 2001-2014. *JAMA 315*(16), 1750–1766.

Chiang, C. L. (1960). A stochastic study of the life table and its applications: I. Probability distributions of the biometric functions. *Biometrics 16*(4), 618–635.

Clark, S. J. (2016). A General Age-Specific Mortality Model with an Example Indexed by Child or Child/Adult Mortality. *arXiv preprint arXiv:1612.01408*.

Coale, A. J., P. Demeny, and B. Vaughan (1966). *Regional Model Life Tables and Stable Populations: Studies in Population*. Elsevier.

Congdon, P. (2009). Life expectancies for small areas: a Bayesian random effects methodology. *International Statistical Review 77*(2), 222–240.

Congdon, P. (2014). *Applied Bayesian modelling*, Volume 595. John Wiley & Sons.

Congdon, P., S. Shouls, and S. Curtis (1997). A multi-level perspective on small-area health and mortality: a case study of England and Wales. *Population, Space and Place 3*(3), 243–263.

Currie, I. D. and M. Durban (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling 2*(4), 333–349.

Currie, I. D., M. Durban, and P. H. C. Eilers (2004). Smoothing and forecasting mortality rates. *Statistical modelling 4*(4), 279–298.

Currie, J. and H. Schwandt (2016). Inequality in mortality decreased among the young while increasing for older adults, 1990–2010. *Science 352*(6286), 708–712.

D'Amato, V., G. Piscopo, and M. Russolillo (2011). The mortality of the Italian population: Smoothing techniques on the Lee-Carter model. *Ann. Appl. Stat. 5*(2A), 705–724.

Daponte, B. O., J. B. Kadane, and L. J. Wolfson (1997). Bayesian demography: projecting the Iraqi Kurdish population, 1977–1990. *Journal of the American Statistical Association 92*(440), 1256–1267.

Deaton, A. S. and C. Paxson (2001). Mortality, education, income, and inequality among American cohorts. In *Themes in the Economics of Aging*, pp. 129–170. University of Chicago Press.

Delwarde, A., M. Denuit, and P. Eilers (2007). Smoothing the Lee–Carter and Poisson log-bilinear models for mortality forecasting: a penalized log-likelihood approach. *Statistical modelling 7*(1), 29–48.

Dwyer-Lindgren, L., A. Bertozzi-Villa, R. W. Stubbs, C. Morozoff, M. J. Kutz, C. Huynh, R. M. Barber, K. A. Shackelford, J. P. Mackenbach, F. J. Van Lenthe, et al. (2016). US county-level trends in mortality rates for major causes of death, 1980-2014. *JAMA 316*(22), 2385–2401.

Eilers, P. H. C., J. Gampe, B. D. Marx, and R. Rau (2008). Modulation models for seasonal time series and incidence tables. *Statistics in Medicine 27*(17), 3430–3441.

Eilers, P. H. C. and B. D. Marx (1996, 05). Flexible smoothing with B-splines and penalties. *Statist. Sci. 11*(2), 89–121.

Elo, I. T. (2009). Social class differentials in health and mortality: Patterns and explanations in comparative perspective. *Annual Review of Sociology 35*, 553–572.

Ewbank, D. C., J. C. Gomez De Leon, and M. A. Stoto (1983). A reducible four-parameter system of model life tables. *Population Studies 37*(1), 105–127.

Ezzati, M., A. B. Friedman, S. C. Kulkarni, and C. J. Murray (2008). The reversal of fortunes: trends in county mortality and cross-county mortality disparities in the United States. *PLoS medicine 5*(4), e66.

Feehan, D. M. (2017). Testing theories of old-age mortality using model selection techniques. *arXiv preprint arXiv:1707.09433*.

Fletcher, R. (2013). *Practical methods of optimization*. John Wiley & Sons.

Foreman, K. J., R. Lozano, A. D. Lopez, and C. J. Murray (2012). Modeling causes of death: an integrated approach using CODEm. *Population Health Metrics 10*(1), 1.

Frydman, C. and R. Molloy (2011). The compression in top income inequality during the 1940s. Technical report, Working paper, MIT.

Gavrilov, L. A. and N. S. Gavrilova (2011). Mortality measurement at advanced ages: a study of the Social Security Administration Death Master File. *North American Actuarial Journal 15*(3), 432–447.

Gelfand, A. E. and A. F. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*(410), 398–409.

Gelman, A. and D. B. Rubin (1992, 11). Inference from Iterative Simulation Using Multiple Sequences. *Statist. Sci. 7*(4), 457–472.

Girosi, F. and G. King (2007). Understanding the Lee-Carter mortality forecasting method. *Copy at http://gking. harvard. edu/files/lc. pdf*.

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

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality. *Philosophical Transactions 27*, 513–585.

Heligman, L. and J. H. Pollard (1980). The age pattern of mortality. *Journal of the Institute of Actuaries 107*(1), 49–80.

Hendi, A. S. (2015). Trends in US life expectancy gradients: the role of changing educational composition. *International Journal of Epidemiology 44*(3), 946–955.

Hill, M. E. and I. Rosenwaike (2001). The Social Security Administration's Death Master File: the completeness of death reporting at older ages. *Soc. Sec. Bull. 64*, 45.

Himes, C. L., S. H. Preston, and G. A. Condran (1994). A relational model of mortality at older ages in low mortality countries. *Population Studies 48*(2), 269–291.

HMD (2018). Human mortality database. Available at http://www.mortality.org/. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).

Horiuchi, S. and J. R. Wilmoth (1998). Deceleration in the age pattern of mortality at older ages. *Demography 35*(4), 391–412.

Hougaard, P. (2012). *Analysis of multivariate survival data*. Springer Science & Business Media.

Hummer, R. A. and E. M. Hernandez (2013). The effect of educational attainment on adult mortality in the United States. *Population Bulletin 68*(1), 1.

Hummer, R. A. and J. T. Lariscy (2011). Educational attainment and adult mortality. In *International handbook of adult mortality*, pp. 241–261. Springer.

Hummer, R. A., R. G. Rogers, and I. W. Eberstein (1998). Sociodemographic differentials in adult mortality: A review of analytic approaches. *Population and Development Review*, 553–578.

Huntington, J. T., M. Butterfield, J. Fisher, D. Torrent, and M. Bloomston (2013). The Social Security Death Index (SSDI) most accurately reflects true survival for older oncology patients. *American Journal of Cancer Research 3*(5), 518.

INSEE (2015). French department mortality data. Personal communication to Magali Barbieri by the Division des statistiques régionales, locales et urbaines, INSEE.

Jarner, S. F. and E. M. Kryger (2011). Modelling adult mortality in small populations: The SAINT model. *ASTIN Bulletin: The Journal of the IAA 41*(2), 377–418.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association 84*(406), 414–420.

Juster, F. T. and R. Suzman (1995). An overview of the Health and Retirement Study. *Journal of Human Resources*, S7–S56.

Kannisto, V. (1988). On the survival of centenarians and the span of life. *Population DStudies 42*(3), 389–406.

Kindig, D. A. and E. R. Cheng (2013). Even as mortality fell in most US counties, female mortality nonetheless rose in 42.8 percent of counties from 1992 to 2006. *Health Affairs 32*(3), 451–458.

King, G. and S. Soneji (2011). The future of death in America. *Demographic Research 25*, 1.

Kochanek, K., S. Murphy, J. Xu, and E. Arias (2017). Mortality in the United States, 2016. Technical Report NCHS Data Brief, no 293., National Center for Health Statistics, Hyattsville, MD.

Krueger, P. M., M. K. Tran, R. A. Hummer, and V. W. Chang (2015). Mortality attributable to low levels of education in the United States. *PloS one 10*(7), e0131809.

Kulkarni, S. C., A. Levin-Rector, M. Ezzati, and C. J. Murray (2011). Falling behind: life expectancy in US counties from 2000 to 2007 in an international context. *Population Health Metrics 9*(1), 16.

Lawn, J. E., S. Cousens, and J. Zupan (2004, 2017/07/28). 4 million neonatal deaths: When? Where? Why? *The Lancet 365*(9462), 891–900.

Lee, R. (2015). The Lee-Carter Model: an update and some extensions. In *Longevity 11 Conference, Lyon, France (September 7 th, 2015)*.

Lee, R. and L. R. Carter (1992). Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association 87*(419), 659–671.

Lozano, R., H. Wang, K. J. Foreman, J. K. Rajaratnam, M. Naghavi, J. R. Marcus, L. Dwyer-Lindgren, K. T. Lofgren, D. Phillips, C. Atkinson, et al. (2011). Progress towards Millennium Development Goals 4 and 5 on maternal and child mortality: an updated systematic analysis. *The Lancet 378*(9797), 1139–1165.

Macintyre, S., A. Ellaway, and S. Cummins (2002). Place effects on health: how can we conceptualise, operationalise and measure them? *Social Science & Medicine 55*(1), 125–139.

Mahy, M. (2003). Measuring child mortality in AIDS-affected countries. Paper presented at the workshop on HIV/AIDS and adult mortality in Developing countries. Available at: `http://www.un.org/esa/population/publications/adultmort/UNICEF_Paper15.pdf`.

Makeham, W. M. (1860). On the law of mortality and construction of annuity tables. *Journal of the Institute of Actuaries 8*(6), 301–310.

Masters, R. K., R. A. Hummer, and D. A. Powers (2012). Educational differences in US adult mortality: A cohort perspective. *American Sociological Review 77*(4), 548–572.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

Missov, T. I., A. Lenart, L. Nemeth, V. Canudas-Romo, and J. W. Vaupel (2015). The Gompertz force of mortality in terms of the modal age at death. *Demographic Research 32*(36), 1031–1048.

Murray, C. J., S. C. Kulkarni, C. Michaud, N. Tomijima, M. T. Bulzacchelli, T. J. Iandiorio, and M. Ezzati (2006). Eight Americas: investigating mortality disparities across races, counties, and race-counties in the United States. *PLoS medicine 3*(9), e260.

National Archives (2018). 1940 Census. Available at `https://1940census.archives.gov/`.

NCHS (2005). National Health Interview Survey Linked Mortality Files. Available at `https://www.cdc.gov/nchs/data-linkage/mortality.htm`.

Nelson, W. B. (2005). *Applied life data analysis*, Volume 577. John Wiley & Sons.

Oestergaard, M. Z., M. Inoue, S. Yoshida, W. R. Mahanani, F. M. Gore, S. Cousens, J. E. Lawn, C. D. Mathers, on behalf of the United Nations Inter-agency Group for Child Mortality Estimation, and the Child Health Epidemiology Reference Group (2011, 08). Neonatal Mortality Levels for 193 Countries in 2009 with Trends since 1990: A Systematic Analysis of Progress, Projections, and Priorities. *PLOS Medicine 8*(8), 1–13.

Ouellette, N. and R. Bourbeau (2011). Changes in the age-at-death distribution in four low mortality countries: A nonparametric approach. *Demographic Research 25*, 595–628.

Paccaud, F., C. S. Pinto, A. Marazzi, and J. Mili (1998). Age at death and rectangularisation of the survival curve: trends in Switzerland, 1969-1994. *Journal of Epidemiology & Community Health 52*(7), 412–415.

Pedersen, J. and J. Liu (2012, 08). Child Mortality Estimation: Appropriate Time Periods for Child Mortality Estimates from Full Birth Histories. *PLoS Medicine 9*(8), e1001289.

Pickett, K. E. and R. G. Wilkinson (2015). Income inequality and health: a causal review. *Social Science & Medicine 128*, 316–326.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Plummer, M. (2012). JAGS Version 3.3.0 user manual. *International Agency for Research on Cancer, Lyon, France*.

Preston, S., P. Heuveline, and M. Guillot (2000). *Demography: measuring and modeling population processes*. Wiley-Blackwell.

Preston, S. H. (1975). The changing relation between mortality and level of economic development. *Population studies 29*(2), 231–248.

Preston, S. H. and I. T. Elo (1999). Effects of age misreporting on mortality estimates at older ages. *Population studies 53*(2), 165–177.

Price, M., J. Klingner, and P. Ball (2013). Preliminary Statistical Analysis of Documentation of Killings in Syria. United Nations Office of the High Commissioner (OHCHR) technical report. Available at: `http://www.ohchr.org/Documents/Countries/SY/PreliminaryStatAnalysisKillingsInSyria.pdf`.

Raftery, A. E., J. L. Chunn, P. Gerland, and H. Ševčíková (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography 50*(3), 777–801.

Raftery, A. E., N. Li, H. Ševčíková, P. Gerland, and G. K. Heilig (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences 109*(35), 13915–13921.

Ruggles, S., C. A. Fitch, P. Kelly Hall, and M. Sobek (2000). IPUMS-USA: Integrated Public Use Microdata Series for the United States. *Handbook of international historical microdata for population research. Minneapolis: Minnesota Population Center*, 259–284.

Saatcioglu, A. and J. L. Rury (2012). Education and the changing metropolitan organization of inequality: A multilevel analysis of secondary attainment in the United States, 1940–1980. *Historical Methods: A Journal of Quantitative and Interdisciplinary History 45*(1), 21–40.

Schmertmann, C., S. M. Cavenaghi, R. M. Assunção, and J. E. Potter (2013). Bayes plus Brass: Estimating total fertility for many small areas from sparse census data. *Population Studies 67*(3), 255–273.

Schmertmann, C., E. Zagheni, J. R. Goldstein, and M. Myrskylä (2014). Bayesian forecasting of cohort fertility. *Journal of the American Statistical Association 109*(506), 500–513.

Sharrow, D. J., S. J. Clark, M. A. Collinson, K. Kahn, and S. M. Tollman (2013). The age pattern of increases in mortality affected by HIV: Bayesian fit of the Heligman-Pollard Model to data from the Agincourt HDSS field site in rural northeast South Africa. *Demographic Research 29*, 1039.

Siler, W. (1983). Parameters of mortality in human populations with widely varying life spans. *Statistics in medicine 2*(3), 373–380.

Sorlie, P. D., E. Backlund, and J. B. Keller (1995). US mortality by economic, demographic, and social characteristics: the National Longitudinal Mortality Study. *American Journal of Public Health 85*(7), 949–956.

Srebotnjak, T., A. H. Mokdad, and C. J. Murray (2010). A novel framework for validating and applying standardized small area measurement strategies. *Population health metrics 8*(1), 26.

Steinsaltz, D. R. and K. W. Wachter (2006). Understanding mortality rate deceleration and heterogeneity. *Mathematical Population Studies 13*(1), 19–37.

Tai, T. H. and A. Noymer (2017). Models for estimating empirical Gompertz mortality: With an application to evolution of the Gompertzian slope. *Population Ecology*, 1–14.

Tuljapurkar, S. and R. D. Edwards (2011). Variance in death and its implications for modeling and forecasting mortality. *Demographic Research 24*, 497.

Ugarte, M., T. Goicoa, and A. Militino (2010). Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics 21*(3-4), 270–289.

United Nations (1982). Model Life Tables for Developing Countries. Technical report, United Nations publication, Sales No. E.81.XIII.7.

United Nations AIDS (UNAIDS) (2015). UNAIDS Gap Report. Available at: `http://www.unaids.org/en/resources/campaigns/2014/2014gapreport/gapreport/`.

United Nations Inter-agency Group for Child Mortality Estimation (IGME) (2015). Levels and Trends in Child Mortality: Report 2015. Available at: `http://childmortality.org/files_v20/download/IGME%20report%202015%20child%20mortality%20final.pdf`.

United Nations (UN) (2015). Millennium Development Goal Progress Report. Available at: `http://www.un.org/millenniumgoals/news.shtml`.

United Nations (UN) (2017). Health - United Nations Sustainable Development. Available at: http://www.un.org/sustainabledevelopment/health/.

US Census Bureau (2014). Population and Housing Unit Estimates. Available at: https://www.census.gov/programs-surveys/popest/data/tables.2014.html.

Vaupel, J. W. and T. I. Missov (2014). Unobserved population heterogeneity. *Demographic Research*.

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing 27*(5), 1413–1432.

Wachter, K. W. (2014). *Essential Demographic Methods*. Harvard University Press.

Waldron, H. (2007). Trends in mortality differentials and life expectancy for male social security-covered workers, by socioeconomic status. *Soc. Sec. Bull. 67*, 1.

Walker, N., K. Hill, and F. Zhao (2012, 08). Child Mortality Estimation: Methods Used to Adjust for Bias due to AIDS in Estimating Trends in Under-Five Mortality. *PLOS Medicine 9*(8), 1–7.

Wilkinson, R. G. (2006). The impact of inequality. *Social Research 73*(2), 711–732.

Willemse, W. and R. Kaas (2007). Rational reconstruction of frailty-based mortality models by a generalisation of Gompertz law of mortality. *Insurance: Mathematics and Economics 40*(3), 468–484.

Wilmoth, J., S. Zureick, V. Canudas-Romo, M. Inoue, and C. Sawyer (2012). A flexible two-dimensional mortality model for use in indirect estimation. *Population Studies 66*(1), 1–28.

Wilmoth, J. R., K. Andreev, D. Jdanov, D. A. Glei, C. Boe, M. Bubenheim, D. Philipov, V. Shkolnikov, and P. Vachon (2007). Methods protocol for the human mortality database. *University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock. URL: http://mortality. org [version 31/05/2007] 9*, 10–11.

Wilmoth, J. R. and S. Horiuchi (1999). Rectangularization revisited: variability of age at death within human populations. *Demography 36*(4), 475–495.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.

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

World Health Organization (WHO) (2013). WHO methods and data sources for global causes of death 2000–2011 (Global Health Estimates Technical Paper WHO/HIS/HSI/GHE/2013.3). Available at: http://www.who.int/healthinfo/statistics/GHE_TR2013-3_COD_MethodsFinal.pdf.

You, D., L. Hug, S. Ejdemyr, P. Idele, D. Hogan, C. Mathers, P. Gerland, J. R. New, and L. Alkema (2015). Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Inter-agency Group for Child Mortality Estimation. *The Lancet 386*(10010), 2275–2286.

Appendix A

Appendices to Chapter 2

A.1 Model summary

The full model is summarized below.

$$
\begin{aligned}
r_{c,i} &\sim N(R_{c,t[c,i]}, \delta_i^2) \\
\delta_i^2 &= \begin{cases} \tau_{c,i}^2 & \text{for VR and SVR data,} \\ \nu_{c,i}^2 + \omega_{s[c,i]}^2 & \text{for non-VR data} \end{cases} \\
R_{c,t} &= f(U_{c,t}) \cdot P_{c,t} \\
\log(f(U_{c,t})) &= \beta_0 + \beta_1 \cdot (\log(U_{c,t}) - \log(\theta))_{[U_{c,t}>\theta]} \\
\log(P_{c,t}) &= \sum_{k=1}^{K_c} B_k(t)\alpha_{c,k} \\
\alpha_{c,k} &= \lambda_c + [\boldsymbol{D}'_{K_c}(\boldsymbol{D}_{K_c}\boldsymbol{D}'_{K_c})^{-1}\boldsymbol{\varepsilon}_c]_k \\
\lambda_c &\sim N(0, \sigma_\lambda^2) \\
\varepsilon_{c,q} &\sim N(0, \sigma_{\varepsilon_c}^2) \\
\log(\sigma_{\varepsilon_c}^2) &\sim N(\chi, \psi^2)
\end{aligned}
$$

where

- $R_{c,t}$ is the true ratio in country c at time t, $R_{c,t} = \frac{N_{c,t}}{U_{c,t}-N_{c,t}}$, where N_{ct} and $U_{c,t}$ are the NMR and U5MR for country c at time t, respectively.
- $r_{c,i}$ is observation i of the ratio in country c.
- $\tau_{c,i}$ is the stochastic standard error, $\nu_{c,i}$ is the sampling error and $\omega_{s[c,i]}^2$ is non-sampling error for series type s.

- β_0 is global intercept; β_1 is global slope with respect to U5MR; θ is the level of U5MR at which β_1 begins to act.
- $P_{c,t}$ is country-specific multiplier for country c at time t.
- $B_k(t)$ is the kth basis spline evaluated at time t and $\alpha_{c,k}$ is splines coefficient k.
- λ_c is the splines intercept for country c.
- $\boldsymbol{D}_{K_c}$ is a $K_c \times (K_c - 1)$ first-order difference matrix: $D_{K_c i,i} = -1$, $D_{K_c i,i+1} = 1$ and $D_{K_c i,j} = 0$ otherwise.
- $\varepsilon_{c,q}$ are fluctuations around the country-specific intercept.
- $\sigma^2_{\varepsilon_c}$ is country-specific smoothing parameter, modeled hierarchically on the log-scale with mean χ and variance ψ^2.

The model was fit in a Bayesian framework. Priors are given by

$$\begin{aligned} \omega &\sim U(0, 40) \\ \beta_0 &\sim N(0, 100) \\ \beta_1 &\sim N(0, 100) \\ \theta &\sim U(0, 500) \\ \sigma_{\lambda c} &\sim U(0, 40) \\ \chi &\sim N(0, 100) \\ \psi &\sim U(0, 40) \end{aligned}$$

A.2 Other aspects of the method

A.2.1 Stochastic errors for VR model

Recall that the observed ratio $r_{c,i}$, which refers to the i-th observation of the ratio in country c, is expressed as a combination of the true ratio and some error, i.e.

$$\begin{aligned} r_{c,i} &= R_{c,t[c,i]} \cdot \epsilon_{c,i} \\ \implies \log(r_{c,i}) &= \log(R_{c,t[c,i]}) + \delta_{c,i} \end{aligned} \tag{A.1}$$

for $c = 1, 2, \ldots, C$ and $i = 1, \ldots, n_c$, where $C = 195$ (the total number of countries) and n_c is the number of observations for country c. The index $t[c, i]$ refers to the observation year for the i-th observation in country c, $\epsilon_{c,i}$ is the error of observation i and $\delta_{c,i} = \log(\varepsilon_{c,i})$.

For VR data series, the error term $\delta_{c,i}$ is modeled as

$$\delta_{c,i} \sim N(0, \tau^2_{c,i}),$$

where $\tau^2_{c,i}$ is the stochastic standard error. These can be obtained once some standard assumptions are made about the distribution of deaths in the first month of life. We assume that deaths below age five d_5 are distributed

$$d_5 \sim Pois(B \times {}_5q_0)$$

where B = live births and ${}_5q_0$ = the probability of death between ages 0 and 5. Additionally, we assume deaths in the first month of life d_n are distributed

$$d_n \sim Bin(d_5, p)$$

where $p = {}_nq_0/{}_5q_0$ and ${}_nq_0$ is the probability of death in the first month of life. Note that the values of ${}_nq_0$ and ${}_5q_0$ come from the raw data.

The stochastic error was obtained via simulation. For each year corresponding to observation i in country c,

- A total of 3,000 simulations of under-five deaths d_5 were drawn from a Poisson distribution $d_5^{(s)} \sim Pois(B \times {}_5q_0)$;
- A total of 3,000 simulations of neonatal deaths d_n were drawn from a Binomial distribution $d_n^{(s)} \sim Bin(d_5^{(s)}, p)$;
- The ratio $y^{(s)} = logit\left(\frac{d_n^{(s)}}{d_5^{(s)}}\right)$ was calculated for each of the simulated samples and the standard error $\tau_{c,i}$ was calculated as $\sigma(\boldsymbol{Y})$ where $\boldsymbol{Y} = (y^{(1)}, y^{(2)}, ...y^{(s)})$, $s = 3,000$.

It is possible that the stochastic variation as assessed in this simulation approach underestimates the true stochastic uncertainty. We used the results from the validation exercise described in Section 3.4.5 to determine the coverage at 80%, 90% and 95% uncertainty levels for VR data only. At all levels the actual coverage level was at least as big as the nominal coverage level. This suggests that either (i) the simulation setup was sufficient to capture stochastic variation in the neonatal deaths and deaths below age 5, or (ii) that any underestimation of stochastic uncertainty is compensated by an overestimation of uncertainty associated with the true ratio. Hence, the validation suggests that credible intervals for the ratio are either well-calibrated or conservative, which is preferable to underestimating the uncertainty associated with the true outcome.

SVR data

For SVR data, the value for the sampling error was imputed based on the sampling error for $U_{c,t[c,i]}$ SVR data, and the observed ratio between the stochastic error of

$r_{c,i}$ and the stochastic error of $U_{c,t[c,i]}$. On average, the stochastic error of $r_{c,i}$ was twice as large as the stochastic error of $U_{c,t[c,i]}$. In addition, the sampling error for $U_{c,t[c,i]}$ SVR data was assumed to be 10%. As such a value of 20% was imputed for the sampling error for $r_{c,i}$ SVR data.

A.2.2 Projection

When producing NMR estimates, generally data are not available up to the most recent year of interest for the majority of countries, and countries may have longer series of missing data. As such, country trajectories needed to be projected forward to the year 2015.

The parameters β_0, β_1 and θ, which make up the expected relation with $U_{c,t}$, are fixed over time, as is the country-specific intercept, λ_c. The component that needs to be projected is the random fluctuations part. These $\varepsilon_{c,k}$ were assumed to be normally distributed around zero, with some variance $\sigma^2_{\varepsilon_c}$ (Equation 3.3). This assumption is used to project the $\varepsilon_{c,k}$'s (and thus the splines).

Start at the first $\alpha_{c,k}$ that is past the last year of observed data. For each time period to be projected:

- Draw $\varepsilon_{c,k} \sim N(0, \sigma^2_{\varepsilon_c})$ to obtain $\alpha_{c,k} = \varepsilon_{c,k} + \alpha_{c,k-1}$
- Repeat to generate α_k for k up to K_c, where K_c is knot number needed to cover the period up to 2015.

The simulated $\varepsilon_{c,k}$ are generally close to zero, so the method essentially propagates the level of the most recent $\alpha_{c,k}$ that overlaps with the data period with the slope of the expected trajectory, as determined by $f(U_{c,t})$. The projection exercise is necessary in order to maintain a consistent level of uncertainty in the estimates.

A.2.3 Recalculation of VR data for small countries

There are several island nations and other small countries that have vital registration data available to calculate NMR. However, observations from these small countries are prone to large stochastic error, which can create erratic trends in NMR over time.

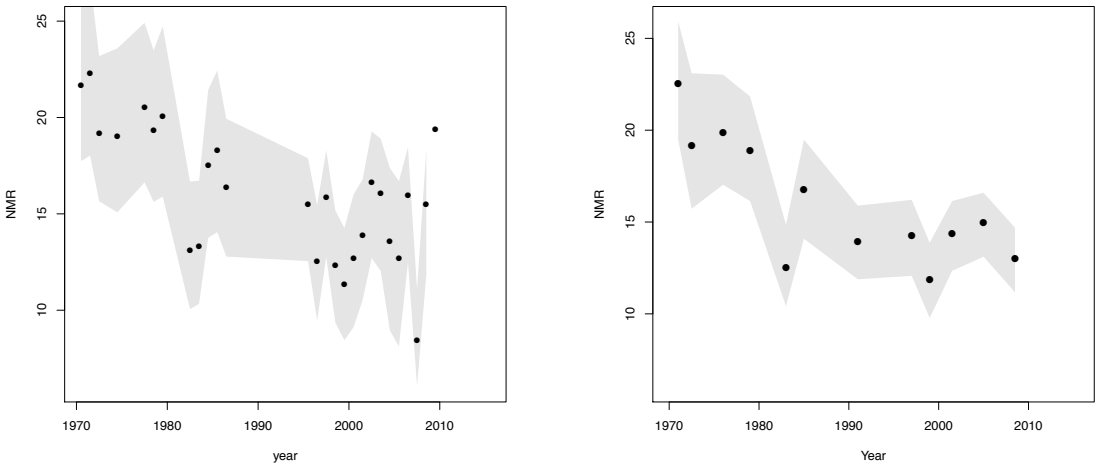
To help avoid this issue, observations from adjacent time periods in a particular country are recombined if the coefficient of variation of the observation is greater than 10%. The result is a smaller set of observations with smaller standard errors

which display a smoother trend. Figure A.1 shows the example of Saint Vincent and the Grenadines on which this process was applied.

NMR is recalculated using the original NMR observations and annual number of live births. For two adjacent years that are to be recalculated:

- The number neonatal deaths in each year is first calculated as NMR $\times$ live births.
- The combined NMR for the two years is then total neonatal deaths divided by total births over the two years.
- The standard error of the new NMR estimate is then recalculated based on the process described in 3.3.4.

After recalculation, the coefficient of variation is calculated for the new estimate. If it is still $> 10\%$, the NMR is recalculated again, recombining with the previous adjacent year.



(a) Original data (b) Recalculated data

Figure A.1: Recalculation of VR data: Saint Vincent and the Grenadines

A.2.4 Crisis deaths

For some countries, there are known natural or political crises that have caused an excess of deaths; for example the Rwandan genocide or, more recently, the Haiti earthquake and conflict in Syria. For the crisis years, the survey data are unlikely to be representative of the actual number of deaths.

Adjustments were made to the relevant crisis country-years, using estimates complied by the World Health Organization (WHO). The WHO uses external data sources on the number of deaths, including the Centre for Research on the Epidemiology of Disasters International Disaster Database (CRED (2012)) and estimates from UN Office of the High Commissioner for Human Rights for the Syrian conflict (Price et al., 2013). The WHO estimates the proportion of deaths that occur under the age of five (WHO (2013)). From there, the best guess of the number of crisis deaths that occur within the first month is simply 1/60th of the total deaths under five years.

Estimation of crisis countries was firstly done without any crisis adjustments. In addition, the global relation with U5MR, $f(U_{c,t})$, is fit to crisis-free $U_{c,t}$ estimates. The relevant adjustments to country-years were then made post estimation. This was to ensure that the crisis deaths, which are specific to particular years, do no have an effect on the splines estimation.

A.2.5 HIV/AIDS countries

Although there have been vast improvements in recent years, many countries in Sub-Saharan Africa still suffer from relatively high levels of HIV/AIDS-related deaths. This has a substantial effect on the child mortality – if children living with HIV are not on antiretroviral treatment, a third will not reach their 1st birthday, and half will not reach their 2nd birthday (UNAIDS (2014)). However, it is unlikely that children with HIV will die within the neonatal period, and so HIV/AIDS itself does not have and explicit effect on the NMR (although there may be indirect effects on mortality, for example through losing their mother to HIV) (Mahy, 2003).

Due to this disproportionate effect of HIV/AIDS on U5MR compared to NMR, there are several adjustments made to the inputs used in the model, which leads to NMR being modeled as a function of 'HIV-free' U5MR. Firstly, the U5MR data used in the ratio observations are adjusted to incorporate reporting bias. This adjustment accounts for the higher maternal mortality among HIV-positive mothers, which leads to under-estimation of U5MR from surveys (Walker et al., 2012). Once adjusted, the AIDS deaths are removed from U5MR, using estimates of deaths provided by UNAIDS (UNAIDS (2014)). The result is a ratio of neonatal to other child mortality which is free of AIDS deaths. In addition, the global relation with U5MR, $f(U_{c,t})$, is fit to AIDS-free U5MR. Unlike the crisis adjustments, no AIDS deaths were added in post-estimation, because it is assumed no neonatal deaths are due to HIV/AIDS.

A.2.6 Countries with no data

There were twelve UN-member countries for which the IGME produces NMR estimates for, but where there are no available data. For these countries, the estimates of NMR are based on the global relation with U5MR, $f(U_{c,t})$. Additionally, some

steps are needed to obtain the appropriate uncertainty around these estimates. For country c:

- Draw $\lambda_c \sim N(0, \sigma_\lambda^2)$;
- Set $\alpha_1 = \lambda_c$;
- Draw $\varepsilon_1 \sim N(0, \sigma_\varepsilon)$; where $\sigma_\varepsilon = e^\chi$ is the global smoothing parameter, based on equation 3.4;
- Set $\alpha_2 = \alpha_1 + \varepsilon_1$;
- Repeat to generate α_k for $k = 3, \ldots, K_c$. K_c is the number of spline knots needed to cover the period 1990–2015.