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

Monica Alexander

1 Introduction[1]

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 uncertainty 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[2] 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

[1] This paper appeared as a chapter of my dissertation, 'Bayesian Methods for Mortality Estimation'.

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

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 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.

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.

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 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,

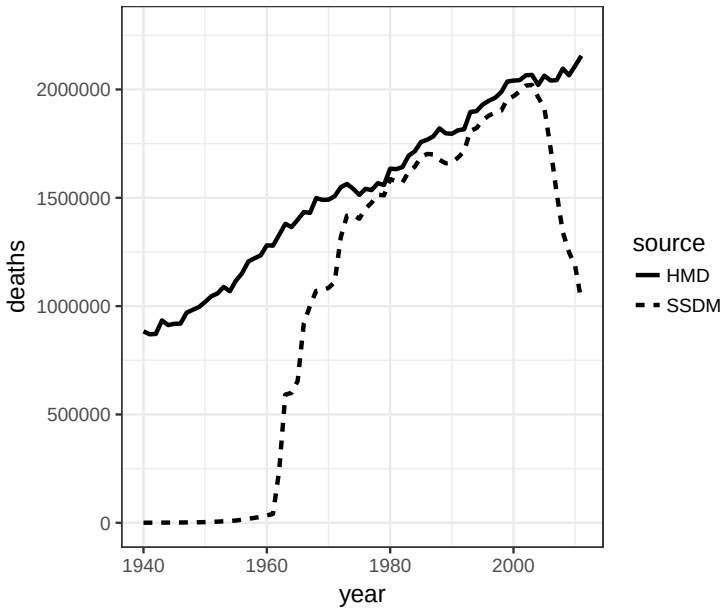

Figure 1: Comparison of the number of deaths at ages 55+ in the SSDM (dotted line) and HMD (solid line), 1940-2011. While deaths in the earlier and later periods are underreported in the SSDM, the period 1975-2005 has close to full coverage.

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. 1 and 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.

2.2 Data preparation

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

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

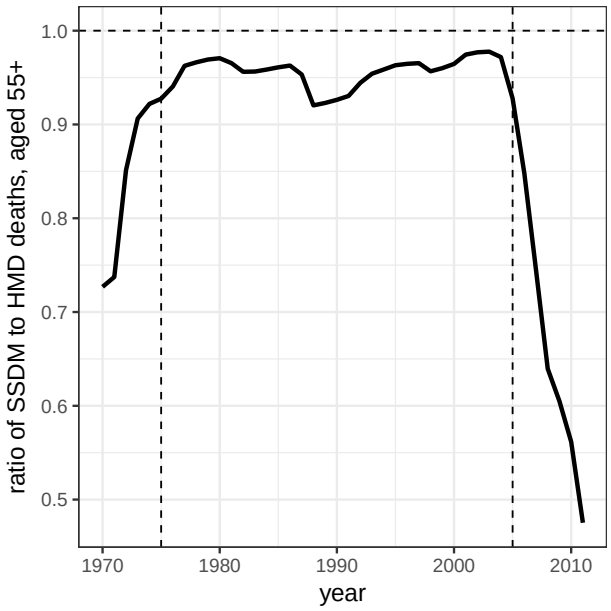

Figure 2: Ratio of deaths at ages 55+ in the SSDM to HMD, 1970-2011. The ratio is around 95% over the period 1975-2005.

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.

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.

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 1: CenSoc match rates by age group

Census age	Match rate (%)	Unique match rate (%)
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. 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 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.

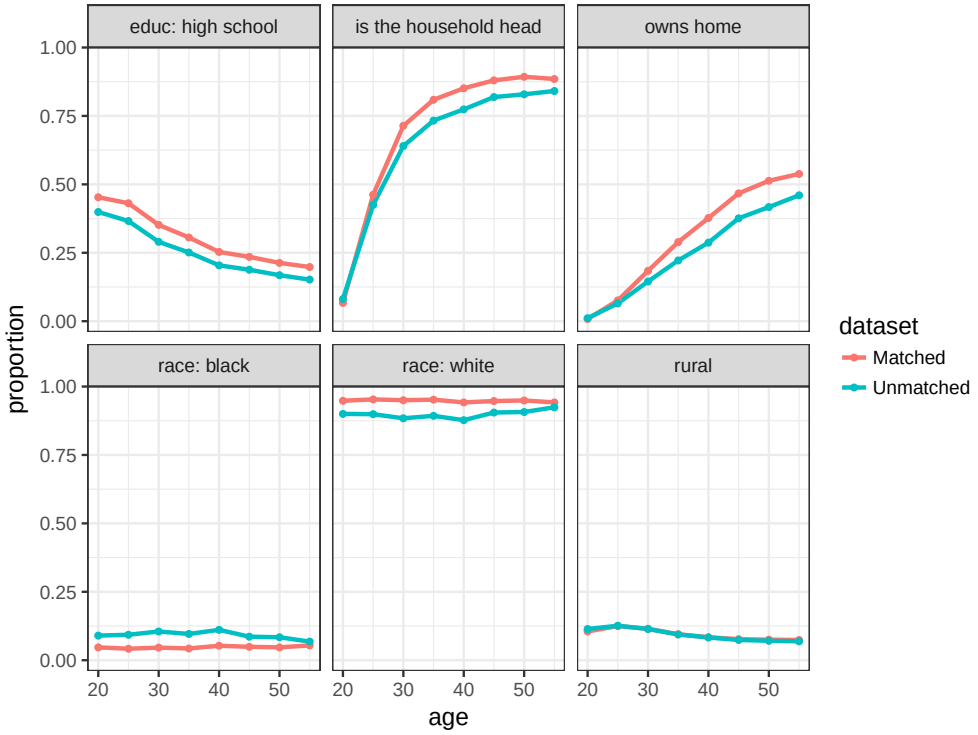

Figure 3: Comparison of socioeconomic characteristics of matched and unmatched datasets. The red line is the proportion of each age group with that characteristics in the CenSoc dataset. The blue line is the same proportion for unmatched individuals in the 1940 census.

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.

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]).

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), we observe deaths only over the period 1975-2005 (the red shaded area in Fig. 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. 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.

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.

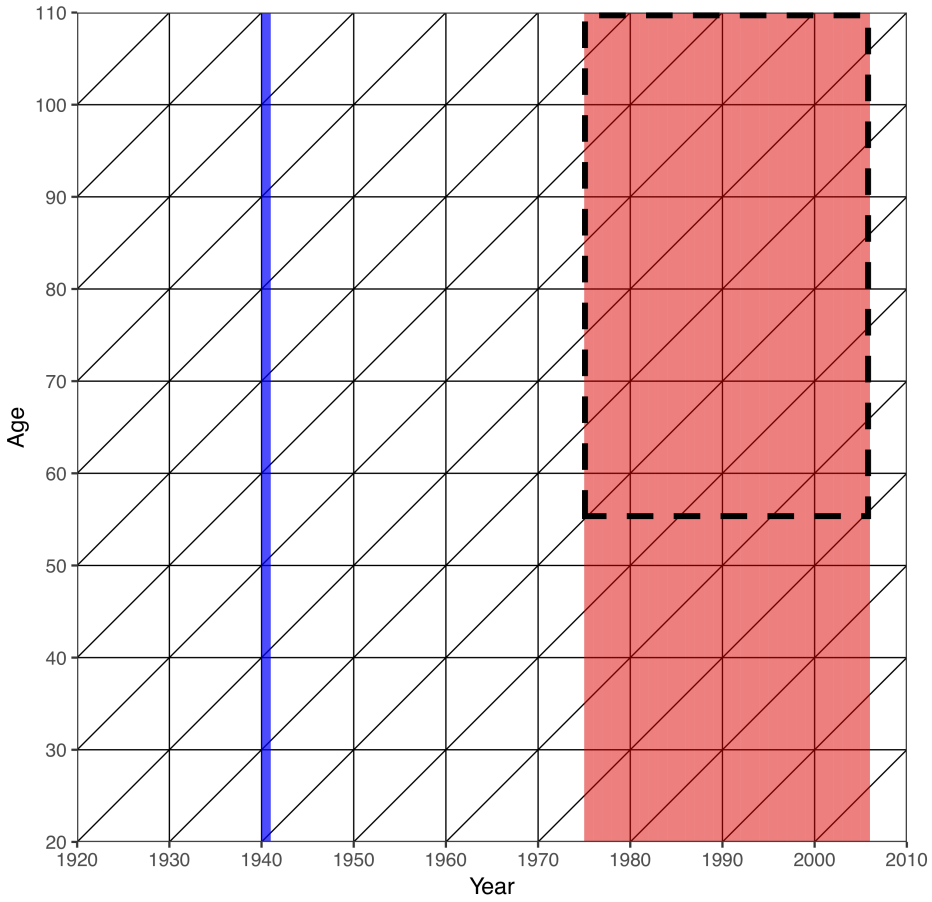

Figure 4: Available information for CenSoc. Socioeconomic information is observed in the 1940 census (blue line); deaths are observed over 1975-2005 in the SSDM (red area). We only consider deaths above age 55, indicated by the dashed box.

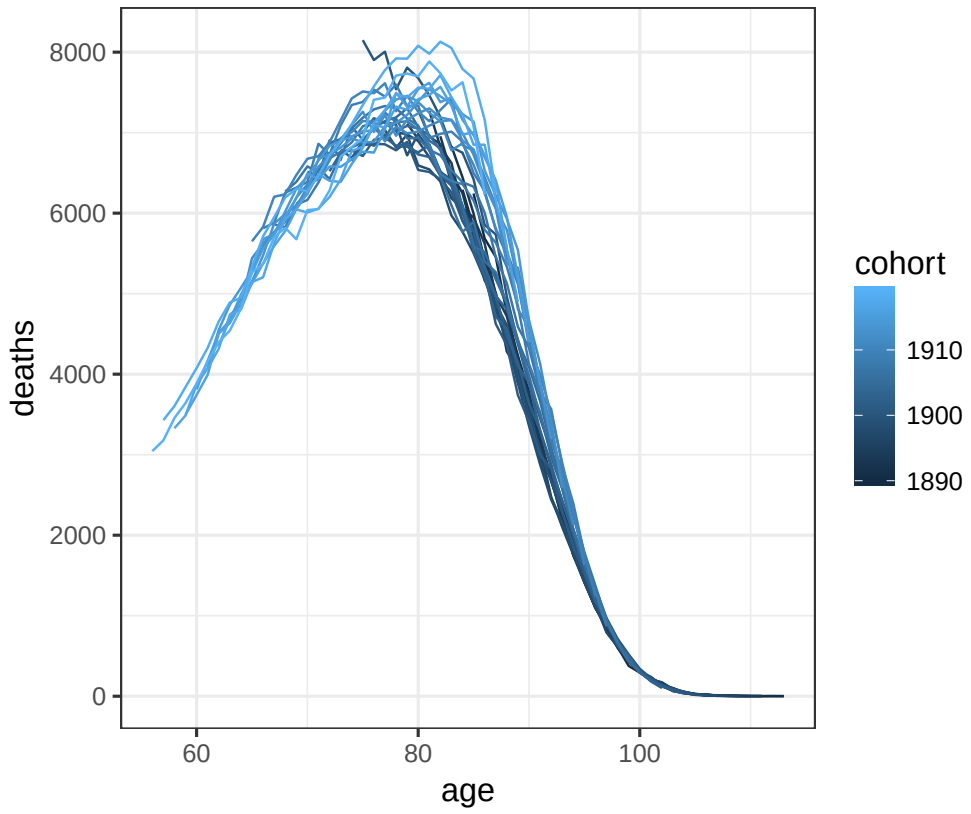

Figure 5: Number of deaths observed by age and cohort in CenSoc. Each line is a different cohort. Every cohort has a different set of observed ages available.

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 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.

4.1 Definition of survival quantities

Define the survivorship function as

$$l(x) = Pr(X > x) \tag{1}$$

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

$$\mu(x) = \lim_{\Delta x \to 0} \frac{Pr(x \leq X < x + \Delta x | x \leq X)}{\Delta X}. \tag{2}$$

This is equivalent to

$$\mu(x) = \frac{-d \log l(x)}{dx}. \tag{3}$$

The cumulative death distribution function is

$$D(x) = 1 - l(x) \tag{4}$$

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

$$d(x) = \frac{-dl(x)}{dx} = \mu(x) l(x) = \mu(x) \exp(-M(x)) \tag{5}$$

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. 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.

4.2 Accounting for truncation

Eq. 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:

$$d^*(x) = \frac{d(x)}{l(x_L) - l(x_U)}. \tag{6}$$

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]):

$$\mathbf{y}|N \sim Multinomial(N, \mathbf{d}^*) \tag{7}$$

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.:

$$D \sim Poisson(N). \tag{8}$$

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

$$y_x \sim Poisson(\lambda_x) \tag{9}$$

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:

$$P(\mathbf{y}|\lambda(\theta)) = \exp\left(-\sum_i \lambda_i\right) \frac{\prod_i \lambda_i^{y_i}}{\prod_i y_i!} \tag{10}$$

with corresponding log-likelihood

$$l(\mathbf{y}|\lambda(\theta)) = -\sum_i \lambda_i + \sum_i y_i \lambda_i - \log \prod_i y_i!. \tag{11}$$

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]).[3] 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,

$$P(\lambda(\theta)|y) = \frac{P(y|\lambda(\theta)) \cdot P(\lambda(\theta))}{P(y)} \tag{12}$$

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 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.

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

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

$$\mu(x) = \alpha e^{\beta x}. \tag{13}$$

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:

$$\log \mu(x) = \alpha + \beta x \tag{14}$$

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. 3, the expression for the Gompertzian survivorship function is

$$l(x) = \exp\left(-\frac{\alpha}{\beta}\left(\exp(\beta x) - 1\right)\right) \tag{15}$$

and it follows from Eq. 5 that the density of deaths at age x, $d(x)$ is

$$d(x) = \mu(x) l(x) = \alpha \exp(\beta x) \exp\left(-\frac{\alpha}{\beta}\left(\exp(\beta x) - 1\right)\right). \tag{16}$$

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. 6 shows values of α and β that lead to mode ages of death within a plausible range (see Eq. 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])

$$M = \frac{1}{\beta} \log\left(\frac{\beta}{\alpha}\right). \tag{17}$$

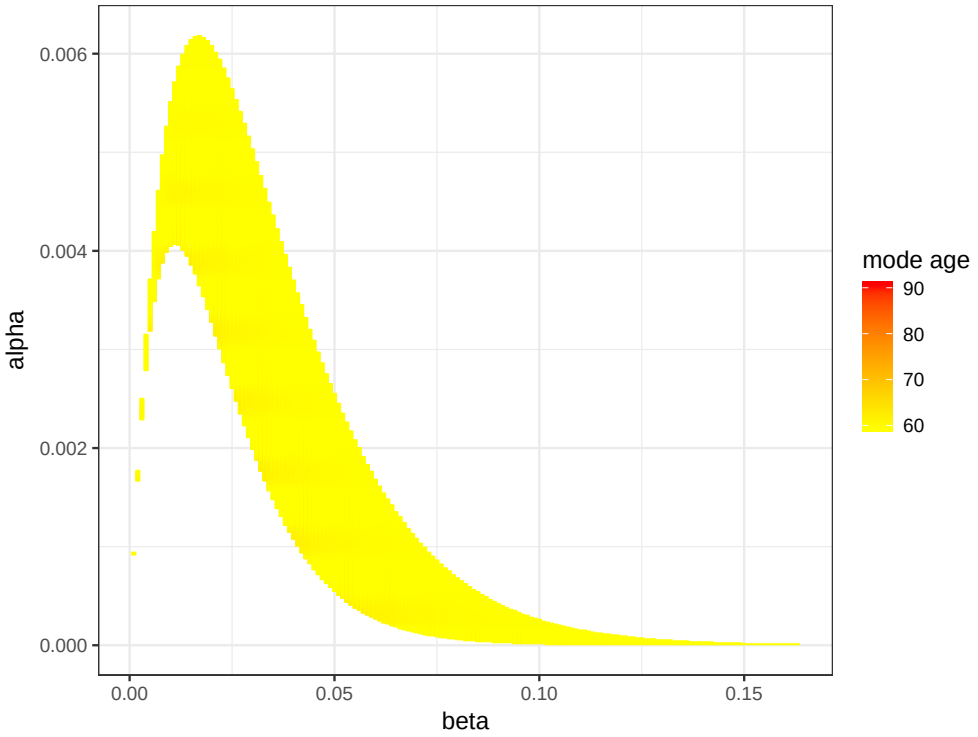

Figure 6: Plausible values of Gompertz parameters α and β given a mode age of between 60-90 years. In general, the larger the value of α, the smaller the value of β. The values of α and β are limited to be below 0.006 and 0.15, respectively.

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

$$\mu(x) = \beta \exp\left(\beta(x - M)\right). \tag{18}$$

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. 5 and 6:

$$\begin{aligned} d^*(x) &= \frac{\mu(x) \cdot l(x)}{l(x_L) - l(x_U)} \\ &= \frac{\beta \exp\left(\beta(x - M)\right) \cdot \exp\left(-\exp\left(-\beta M\right)\left(\exp(\beta x) - 1\right)\right)}{\exp\left(-\exp\left(-\beta M\right)\left(\exp(\beta x_L) - 1\right)\right) - \exp\left(-\exp\left(-\beta M\right)\left(\exp(\beta x_U) - 1\right)\right)} \end{aligned} \tag{19}$$

5.2 Bayesian hierarchical model

Eq. 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. 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 4.3, 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. 9 and 8, the observed deaths by age and cohort are distributed

$$D_c \sim Poisson(N_c) \tag{20}$$

$$y_{c,x} \sim Poisson(\lambda_{c,x}) \tag{21}$$

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

5.2.1 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. 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:

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

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:

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

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. 20, the model for d^*_x will be based on a principal components regression:

$$\text{logit } d_x^* = P_{0,x} + \beta_1 P_{1,x} + \beta_2 P_{2,x} \tag{22}$$

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.

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.

6.1.1 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

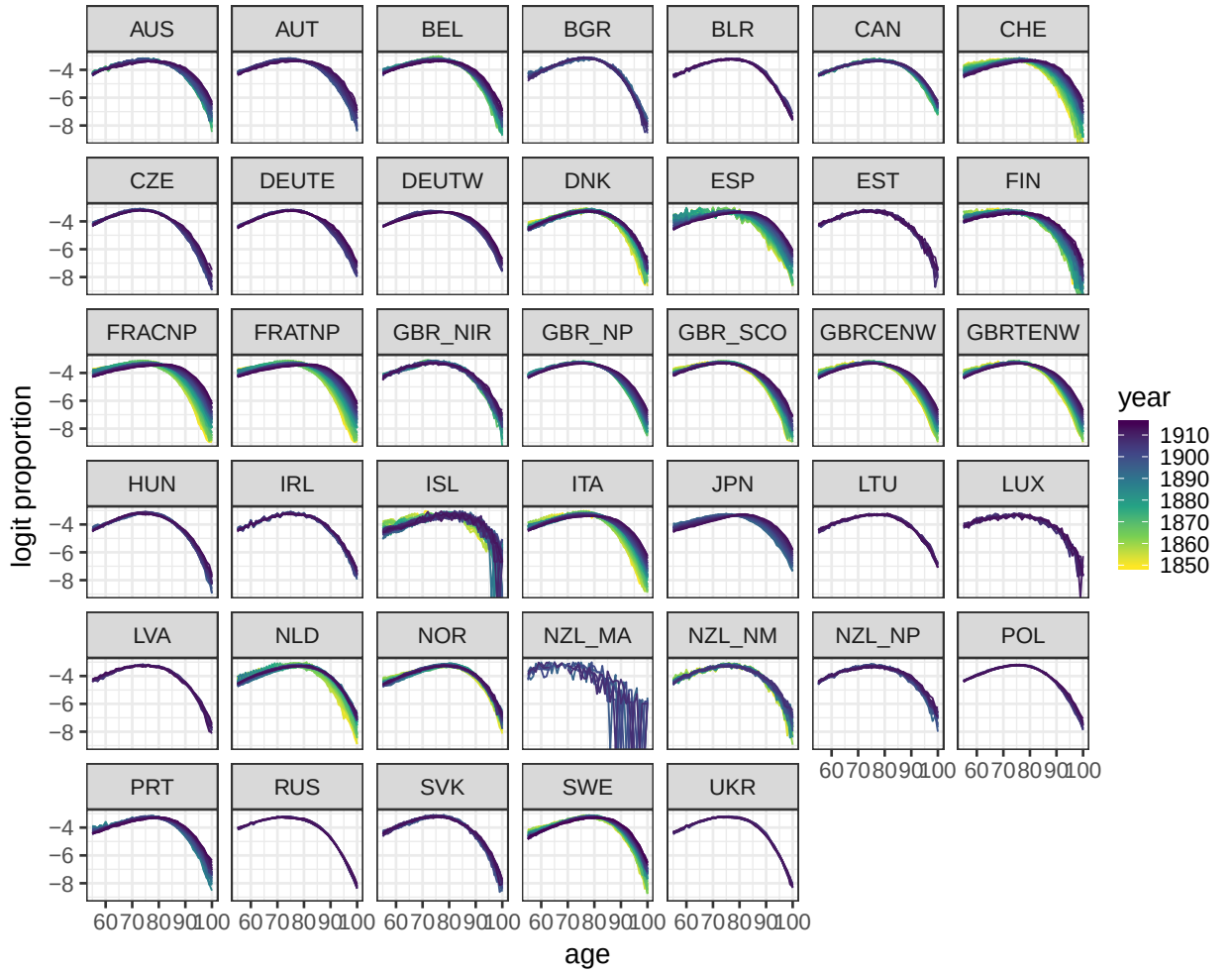

Figure 7: Death distributions in HMD by country and cohort, ages 55-105. The plots show the proportion of deaths at each age, plotted on the logit scale. Each line is a different cohort.

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. 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. 8 shows the mean death distribution and first two principal components obtained from this matrix.[4] The mean schedule gives a shape of baseline mortality across the ages, with mortality peaking 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

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

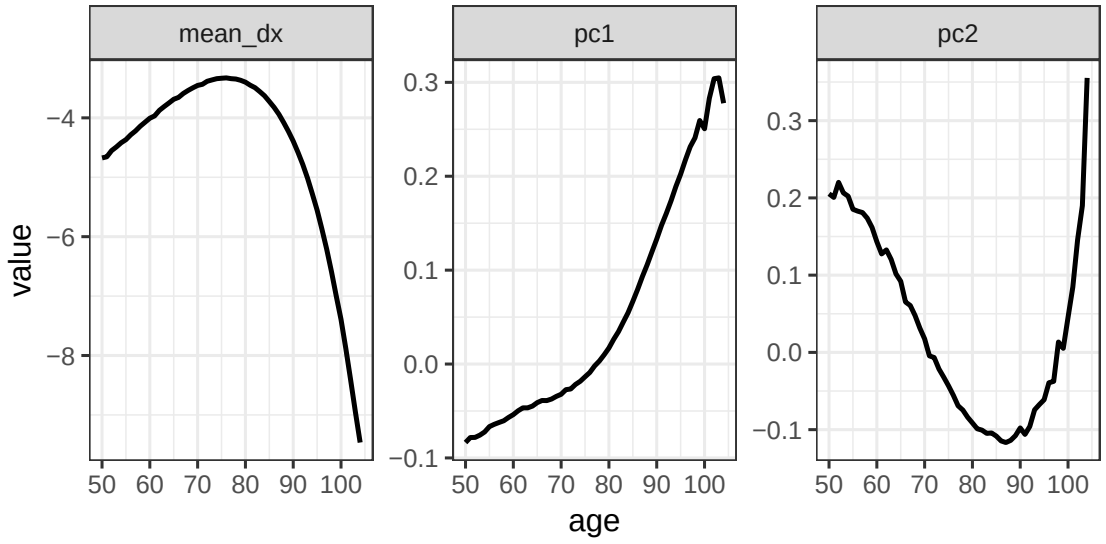

Figure 8: Mean death schedule and first two principal components derived from HMD data shown in Fig. 7. The components are derived from data transformed to be on the logit scale.

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. 9 shows two example death distributions that can be derived from the combination of the curves shown in in Fig. 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.

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

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

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

$$\text{logit } d^*_{c,x} = P_{0,x} + \beta_{1,c} P_{1,x} + \beta_{2,c} P_{2,x} \tag{23}$$

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. 8.

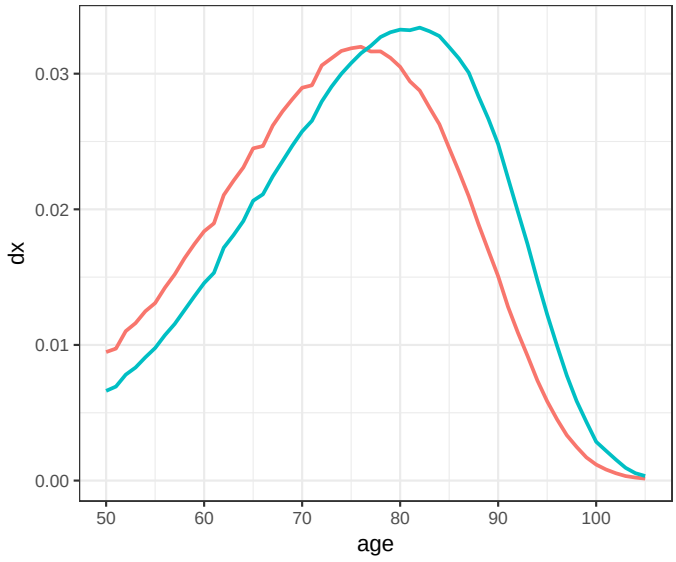

Figure 9: Two example death distributions based on different linear combinations of curves in Fig. 8. For the red curve, the equation is logit $^{-1}(P_{0,x} + 2.5P_{1,x} + 1.25P_{2,x})$. For the blue curve the equation is logit $^{-1}(P_{0,x} + 6P_{1,x} + 0.8P_{2,x})$.

- 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. 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 $\beta_{1,c}$:

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

6.2.1 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:

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

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 available 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.

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. 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.

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.

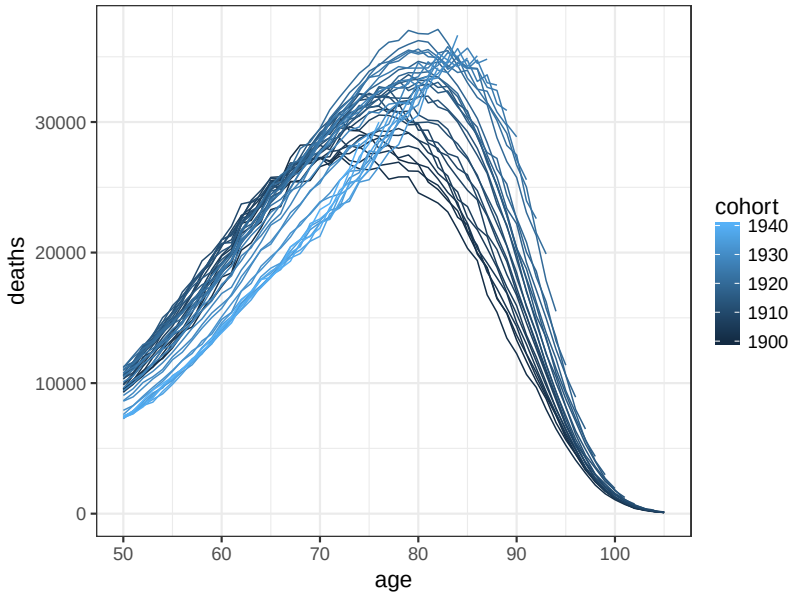

Figure 10: Death counts by age, United States, males, cohorts 1900-1940, ages 50-105. Each line is a different cohort. Data come from the HMD.

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.

7.4 Gompertz results

Results from fitting the truncated Gompertz hierarchical model are shown in Figs. 11-13. The mode age of death is steadily increasing over time, from around age 76 in cohort 1900 to 84 in cohort 1940 (Fig. 11). In

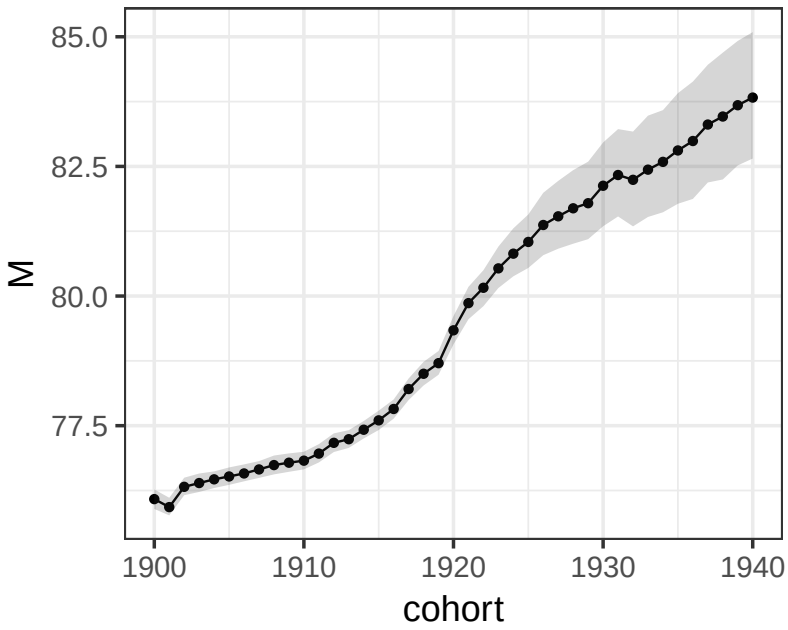

Figure 11: Estimates of Gompertz mode age of death, United States. Median posterior estimates are shown by the dots. The shaded area represents the 95% uncertainty interval.

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. 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.

7.5 Principal component regression results

Results from fitting the principal components regression model are shown in Fig. s 14 and 15. The coefficient on the first principal component steadily increased over cohorts (Fig. 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. 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.

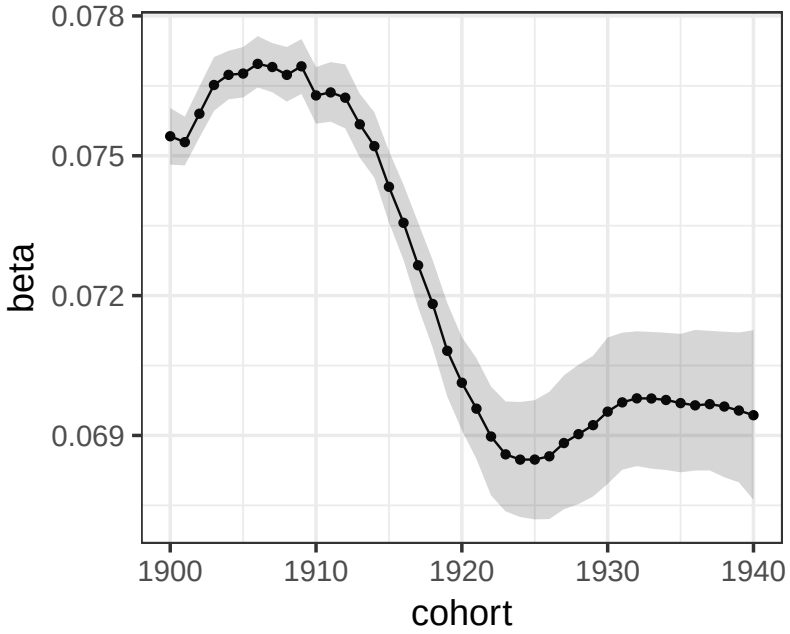

Figure 12: Estimates of Gompertz β, United States. Median posterior estimates are shown by the dots. The shaded area represents the 95% uncertainty interval.

7.6 Comparison of models

Figs. 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.

Fig. 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.

7.6.1 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.

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:

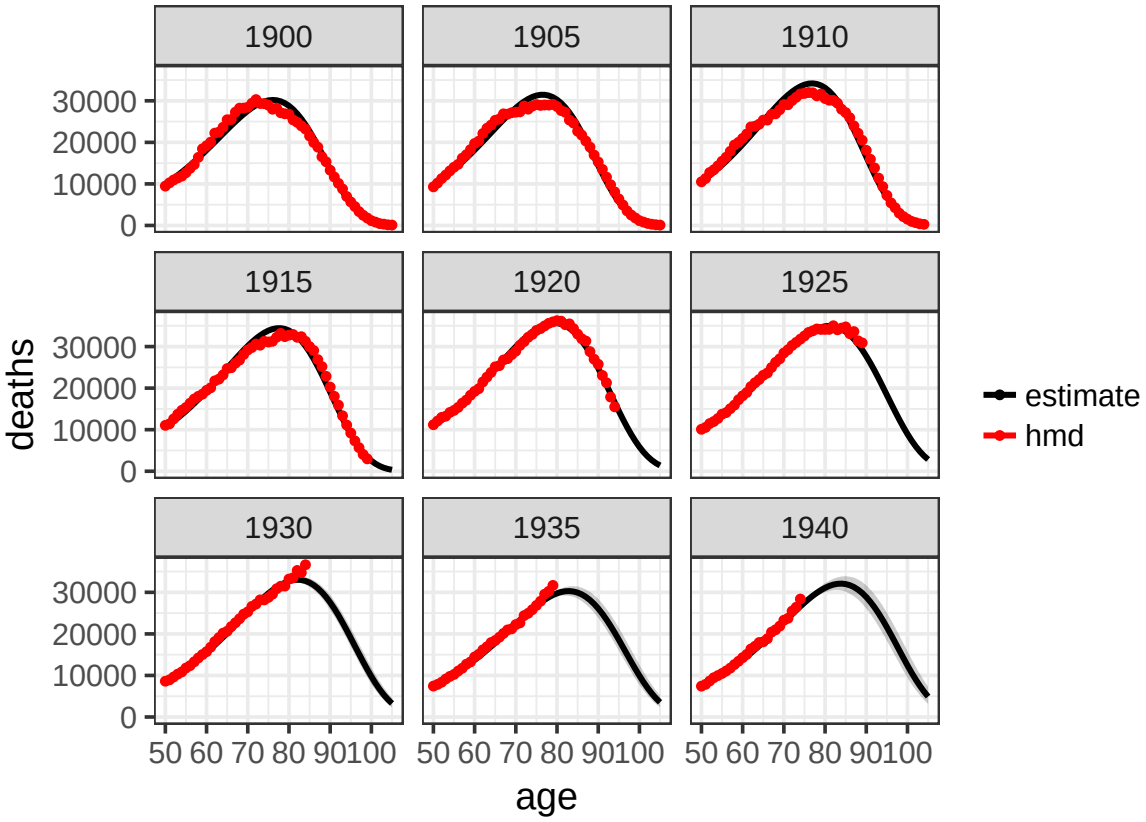

Figure 13: Truncated Gompertz model estimates and HMD data of deaths by age for nine cohorts between 1900 and 1940. Data from HMD are shown by the red dots. The estimates and associated 95% uncertainty intervals are shown by the black lines and shaded areas.

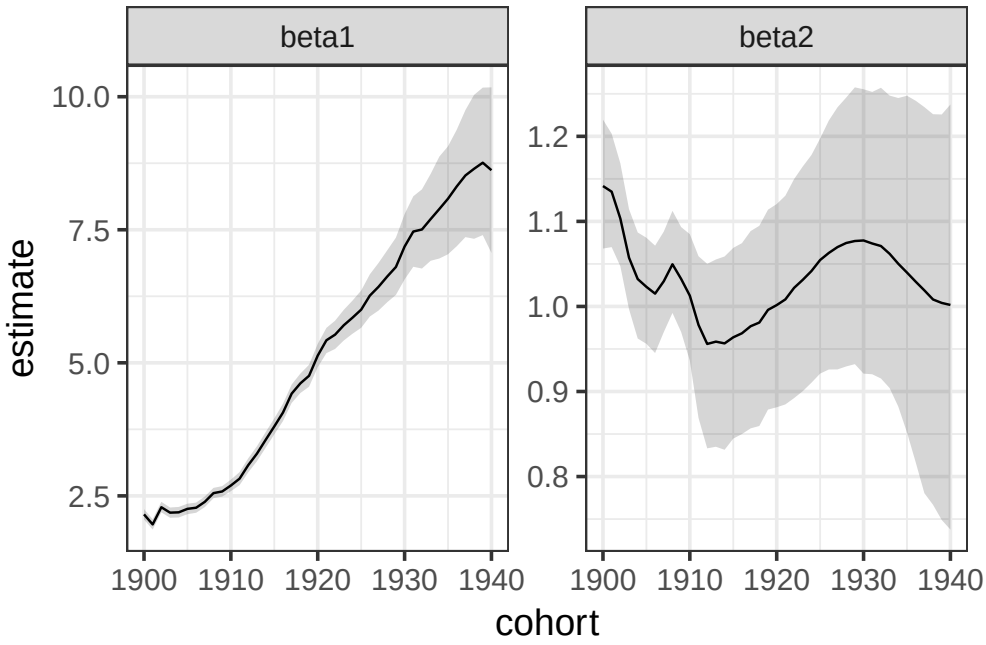

Figure 14: Estimates of principal component coefficients $\beta_{1,c}$ (left) and $\beta_{2,c}$ (right) across cohorts. Median posterior estimates are shown by the black lines. The shaded areas represent the 95% uncertainty intervals.

$$\text{RMSE} = \sqrt{\frac{1}{A}\sum_{x=1}^{A}(\hat{y}_{c,x} - y^*_{c,x})^2}, \tag{24}$$

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

$$\text{RMSE} = \sqrt{\frac{1}{C}\sum_{c=1}^{C}(\hat{y}_{c,x} - y^*_{c,x})^2}, \tag{25}$$

where C is the number of cohorts. Figs. 18 and 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 and the coverage of such intervals assessed. For example, we would expect 95% prediction intervals of $y^{new}_{c,x}$ to include the observed values $y_{c,x}$ 95% of the time.

Fig. 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.

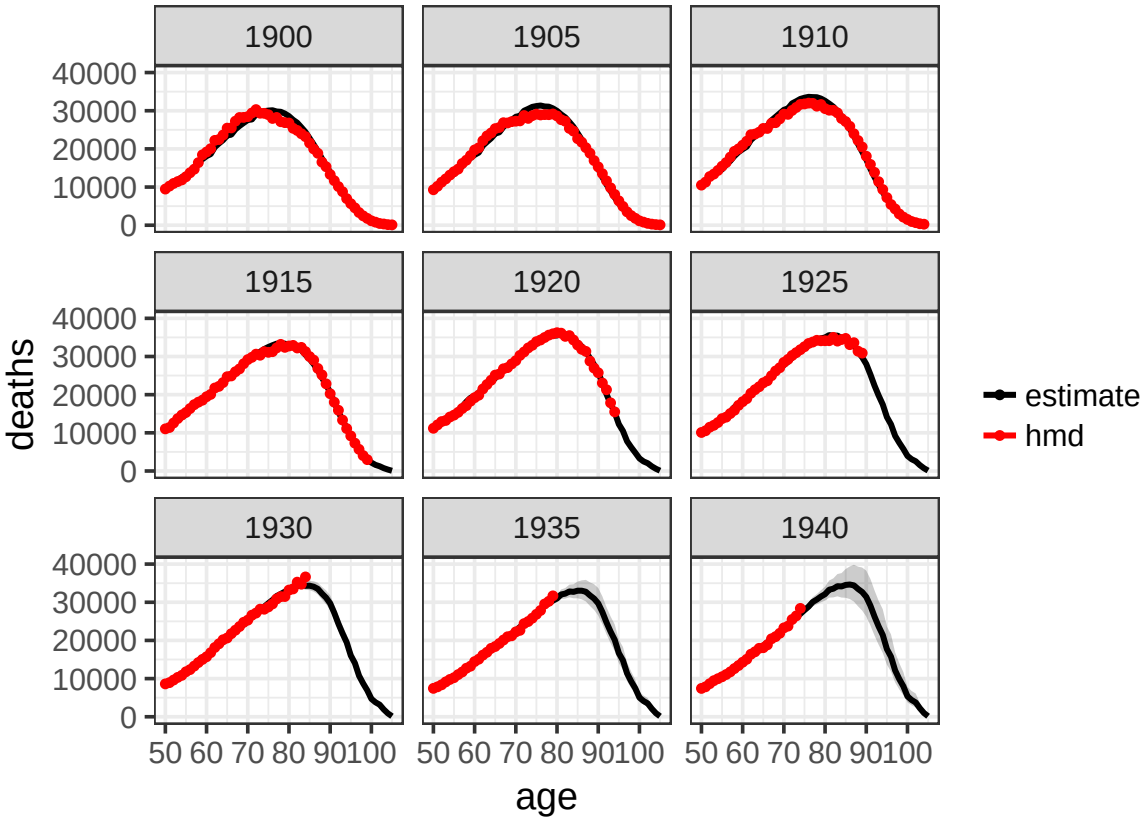

Figure 15: Principal component estimates and HMD data of deaths by age for nine cohorts between 1900 and 1940. Data from HMD are shown by the red dots. The estimates and associated 95% uncertainty intervals are shown by the black lines and shaded areas.

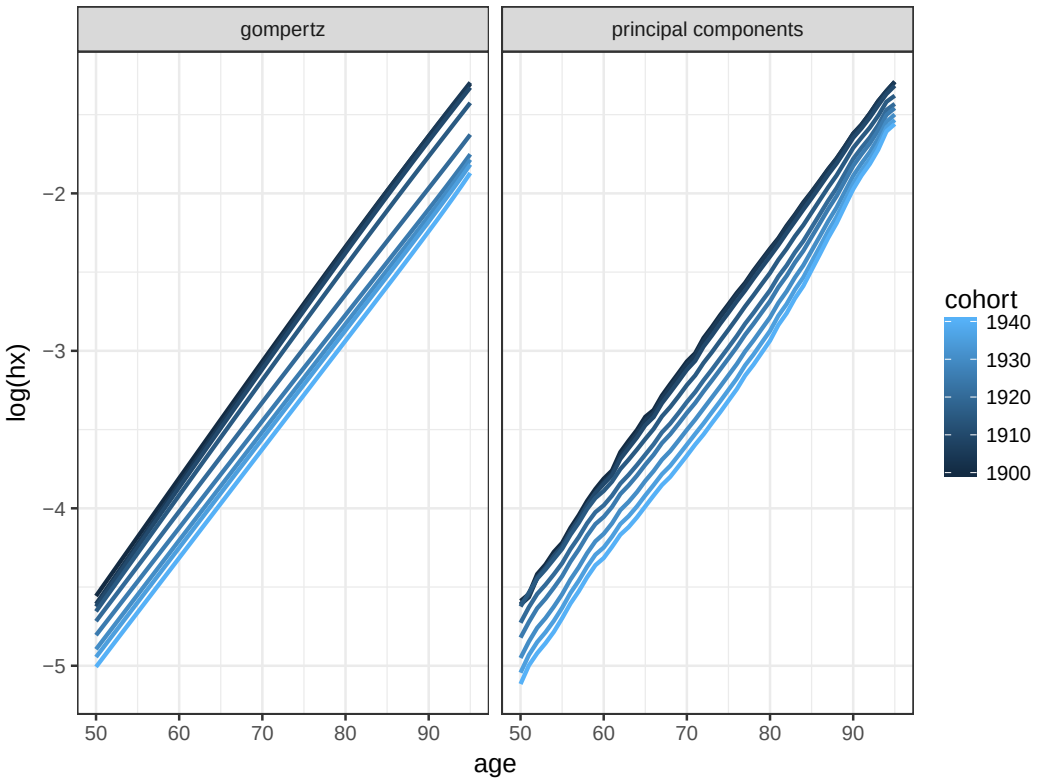

Figure 16: Estimated (log) hazard rates by cohort for the truncated Gompertz (left) and principal components model (right). Each line represents a different cohort.

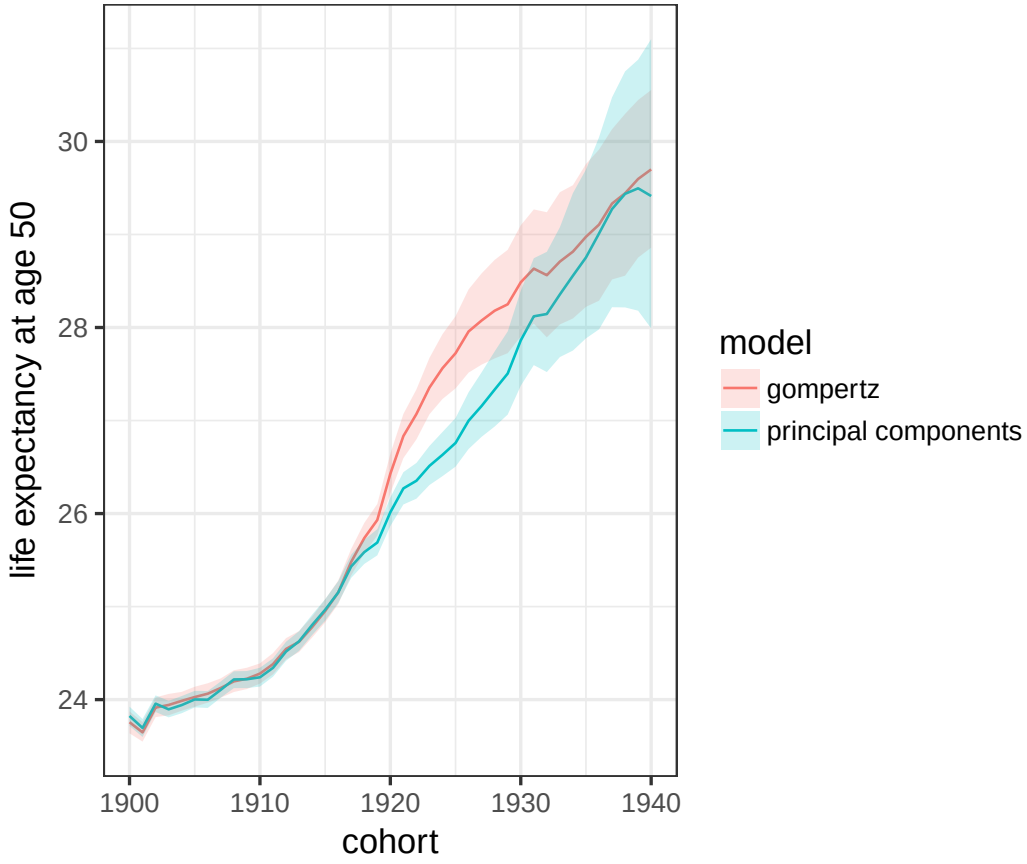

Figure 17: Estimated life expectancy at age 50 by cohort for the Gompertz (red line) and principal components (blue line) models. The median estimates are shown as the lines, and 95% uncertainty intervals are shown by the shaded areas.

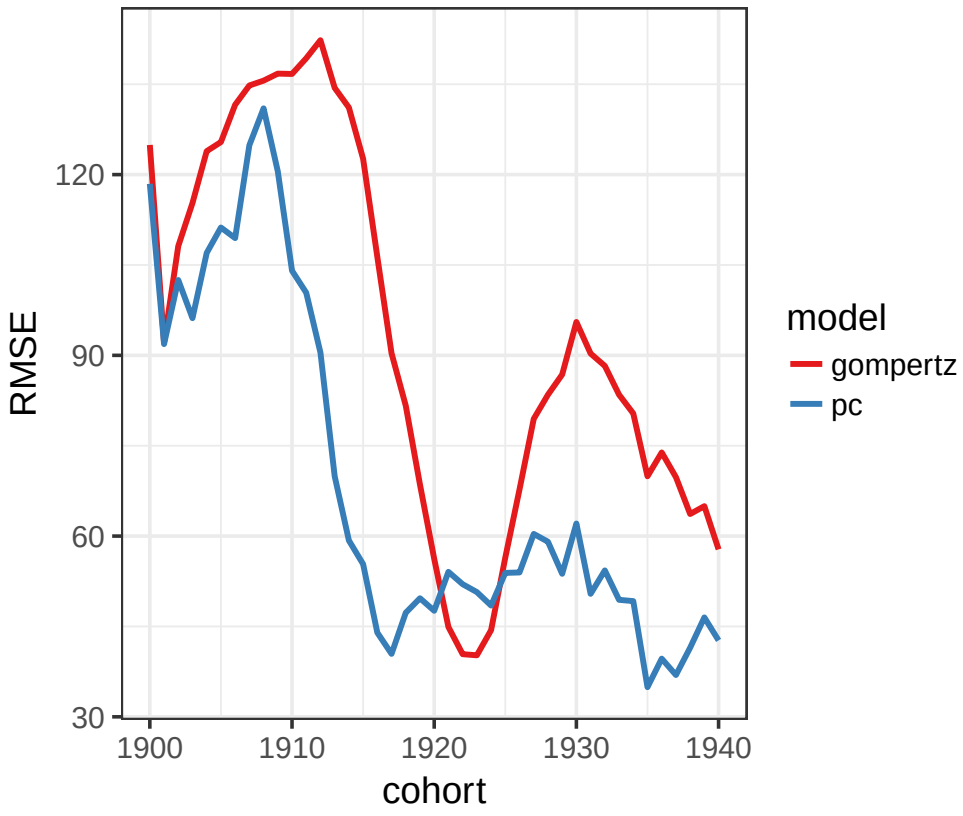

Figure 18: RMSE by cohort for the Gompertz (red line) and principal components (blue line) models.

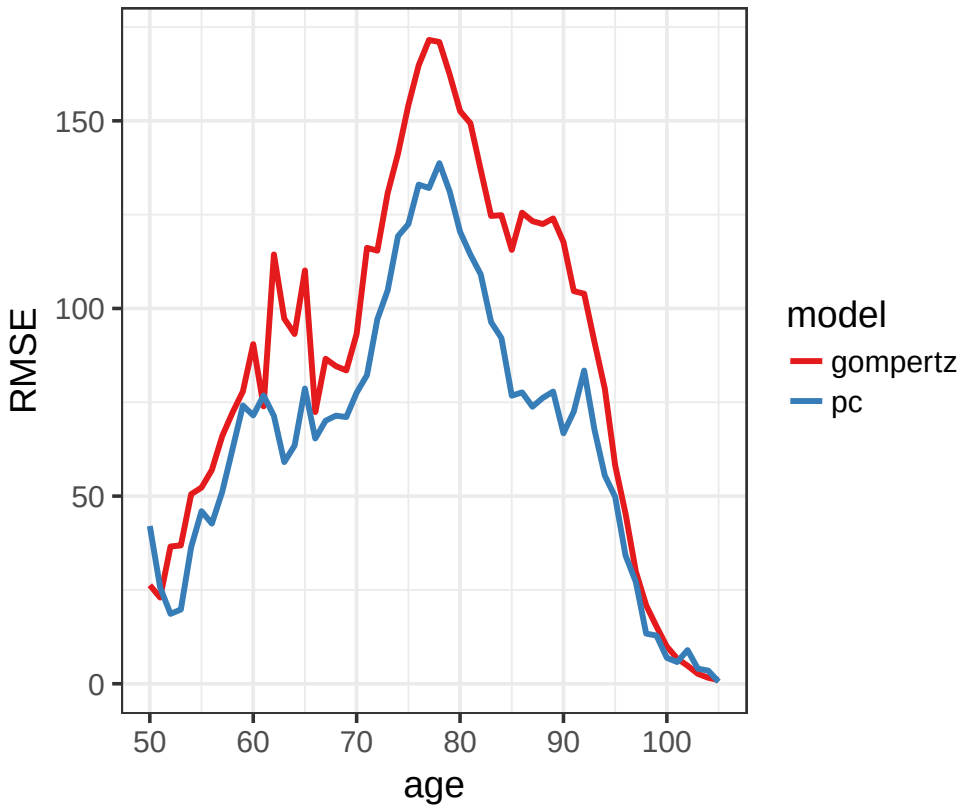

Figure 19: RMSE by age for the Gompertz (red line) and principal components (blue line) models.

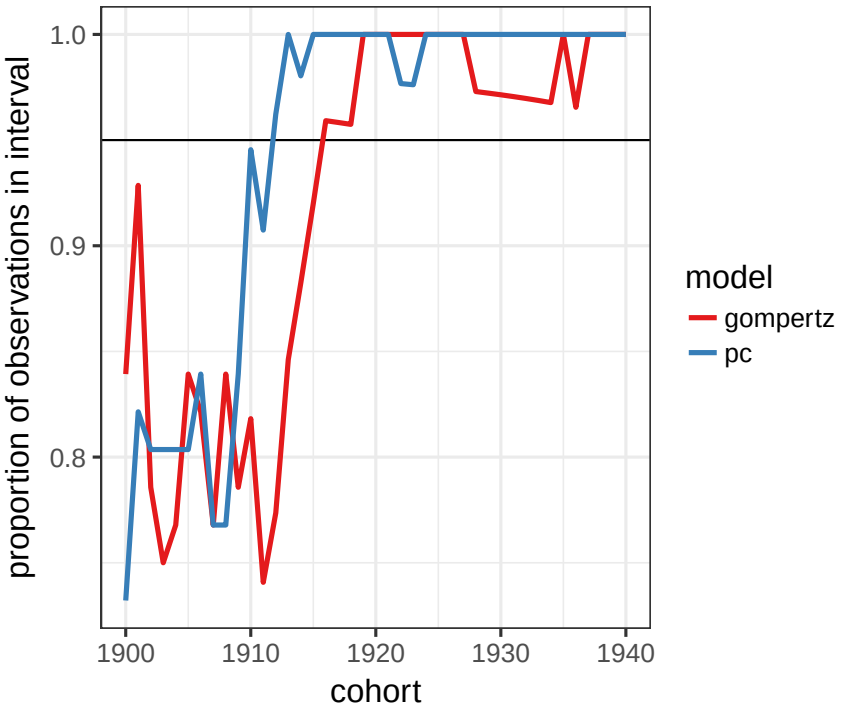

Figure 20: Coverage of 95% prediction intervals for the Gompertz (red line) and principal components (blue line) models. If the uncertainty intervals are well-calibrated, the coverage of the prediction intervals would be expected to be 95%.

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 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.

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.[5] The method can be extended to allow for differing mortality trends by socioeconomic group, as shown below.

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)
- 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:

[5] 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.

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

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. $e_{50,c,g}$.

8.1.1 Results

Fig. 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. 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. 22. In general, mortality disparities between the least and 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. 22 suggests the widening disparity is a consequence of both increases in the higher education groups, and decreases in the lower-educated groups. As illustrated in Fig. 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.

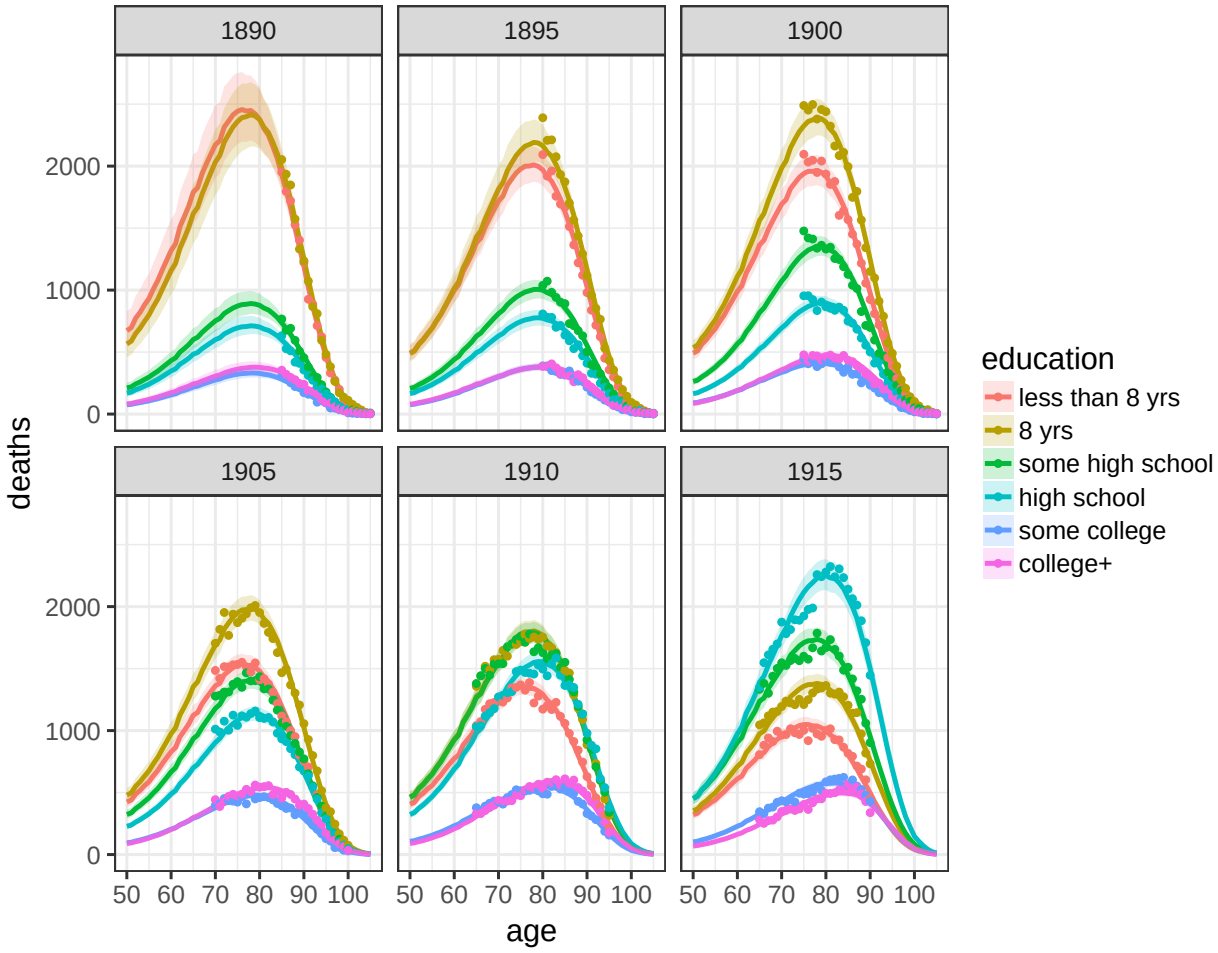

Figure 21: Estimated and observed death counts in CenSoc by education level, cohorts 1890-1915. The available data are shown by the dots, the estimates and associated 95% uncertainty intervals are shown by the lines and shaded areas.

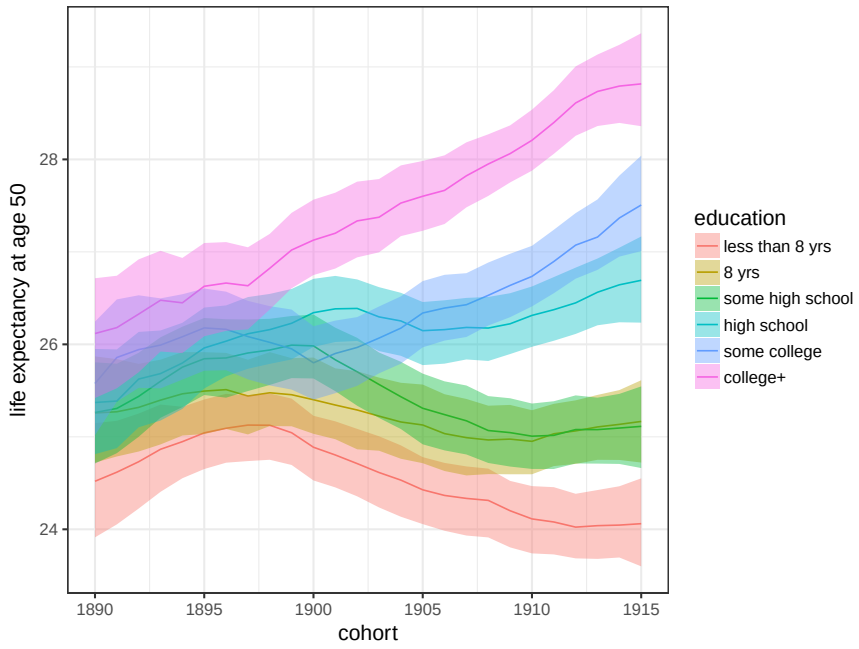

Figure 22: Life expectancy at age 50 by education level, cohorts 1890-1915. The shaded areas represent 95% uncertainty intervals.

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 $d^*_{c,g,x}$.

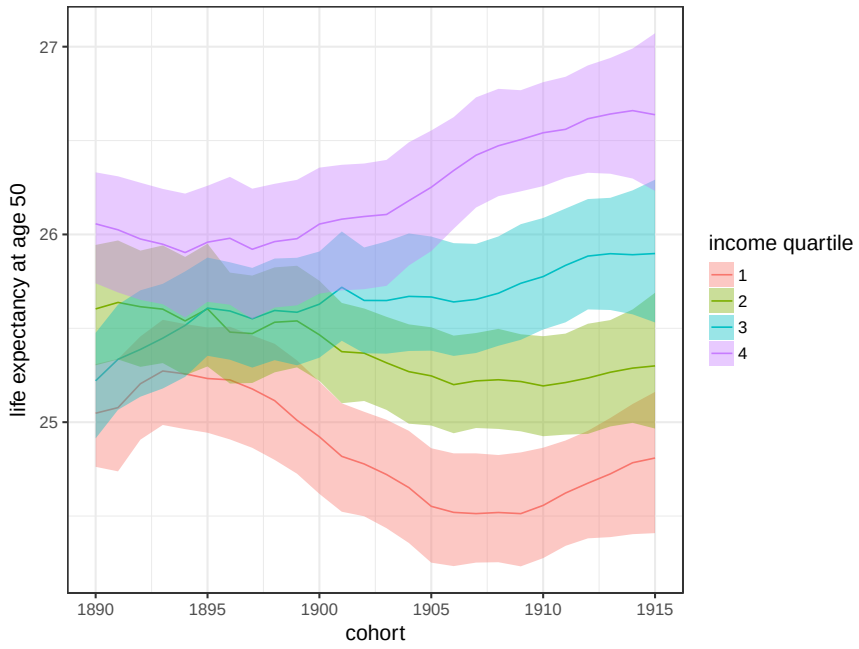

Figure 23: Life expectancy at age 50 by income group, cohorts 1890-1915. The shaded areas represent 95% uncertainty intervals.

8.2.1 Results

Fig. 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 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]).

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. Differences 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

Ran Abramitzky, Roy Mill, and Santiago Pérez. Linking individuals across historical sources: a fully automated approach. Technical report, National Bureau of Economic Research, 2018.

Monica Alexander, Emilio Zagheni, and Magali Barbieri. A flexible bayesian model for estimating subnational mortality. *Demography*, 54(6):2025–2041, 2017.

Rohan Alexander and Zachary Ward. Age at arrival and assimilation during the age of mass migration. *Journal of Economic History*, page Forthcoming, 2018.

Leontine Alkema and Jin Rou New. Global estimation of child mortality using a bayesian b-spline bias-reduction model. *The Annals of Applied Statistics*, 8(4):2122–2149, 2014.

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

Marie-Pier Bergeron-Boucher, Marcus Ebeling, and Vladimir Canudas-Romo. Decomposing changes in life expectancy: Compression versus shifting mortality. *Demographic Research*, 33:391–424, 2015.

Dan A Black, Yu-Chieh Hsu, Seth G Sanders, Lynne Steuerle Schofield, and Lowell J Taylor. The methuselah effect: The pernicious impact of unreported deaths on old-age mortality estimates. *Demography*, 54(6): 2001–2024, 2017.

Barry Bosworth and Kathleen Burke. Differential mortality and retirement benefits in the health and retirement study. *Center for Retirement Research Working Paper*, 2014(4), 2014.

Oskar Burger and Trifon I Missov. Evolutionary theory of ageing and the problem of correlated gompertz parameters. *Journal of Theoretical Biology*, 408:34–41, 2016.

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

Anne Case and Angus Deaton. Mortality and morbidity in the 21st century. *Brookings papers on Economic Activity*, 2017:397, 2017.

Raj Chetty, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. The association between income and life expectancy in the united states, 2001-2014. *JAMA*, 315(16):1750–1766, 2016.

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

Samuel J Clark. A general age-specific mortality model with an example indexed by child or child/adult mortality. *arXiv preprint arXiv:1612.01408*, 2016.

Iain D Currie, Maria Durban, and Paul H C Eilers. Smoothing and forecasting mortality rates. *Statistical modelling*, 4(4):279–298, 2004.

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

Angus S Deaton and Christina Paxson. Mortality, education, income, and inequality among american cohorts. In *Themes in the Economics of Aging*, pages 129–170. University of Chicago Press, 2001.

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

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

Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

Carola Frydman and Raven Molloy. The compression in top income inequality during the 1940s. Technical report, Working paper, MIT, 2011.

Leonid A Gavrilov and Natalia S Gavrilova. Mortality measurement at advanced ages: a study of the social security administration death master file. *North American Actuarial Journal*, 15(3):432–447, 2011.

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

Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statist. Sci.*, 7(4):457–472, 11 1992. doi: 10.1214/ss/1177011136. URL http://dx.doi.org/10.1214/ss/1177011136.

Federico Girosi and Gary King. *Demographic forecasting*. Princeton University Press, 2008.

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

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

Mark E Hill and Ira Rosenwaike. The social security administration's death master file: the completeness of death reporting at older ages. *Soc. Sec. Bull.*, 64:45, 2001.

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

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

Philip Hougaard. *Analysis of multivariate survival data*. Springer Science & Business Media, 2012.

Robert A Hummer and Elaine M Hernandez. The effect of educational attainment on adult mortality in the united states. *Population Bulletin*, 68(1):1, 2013.

Robert A Hummer and Joseph T Lariscy. Educational attainment and adult mortality. In *International handbook of adult mortality*, pages 241–261. Springer, 2011.

Robert A Hummer, Richard G Rogers, and Isaac W Eberstein. Sociodemographic differentials in adult mortality: A review of analytic approaches. *Population and Development Review*, pages 553–578, 1998.

Justin T Huntington, Mathew Butterfield, James Fisher, Daniel Torrent, and Mark Bloomston. The social security death index (ssdi) most accurately reflects true survival for older oncology patients. *American Journal of Cancer Research*, 3(5):518, 2013.

Matthew A Jaro. 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, 1989.

F Thomas Juster and Richard Suzman. An overview of the health and retirement study. *Journal of Human Resources*, pages S7–S56, 1995.

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

KD Kochanek, SL Murphy, JQ Xu, and E Arias. Mortality in the united states, 2016. Technical Report NCHS Data Brief, no 293., National Center for Health Statistics, Hyattsville, MD, 2017.

Patrick M Krueger, Melanie K Tran, Robert A Hummer, and Virginia W Chang. Mortality attributable to low levels of education in the united states. *PloS one*, 10(7):e0131809, 2015.

Ronald Lee and Lawrence R. Carter. Modeling and forecasting u.s. mortality. *Journal of the American Statistical Association*, 87(419):659–671, 1992. ISSN 01621459. URL http://www.jstor.org/stable/2290201.

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

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

P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989. ISBN 9780412317606. URL http://books.google.com/books?id=h9kFH2_FfBkC.

Trifon I Missov, Adam Lenart, Laszlo Nemeth, Vladimir Canudas-Romo, and James Walton Vaupel. The gompertz force of mortality in terms of the modal age at death. *Demographic Research*, 32(36):1031–1048, 2015.

Christopher JL Murray, Sandeep C Kulkarni, Catherine Michaud, Niels Tomijima, Maria T Bulzacchelli, Terrell J Iandiorio, and Majid Ezzati. Eight americas: investigating mortality disparities across races, counties, and race-counties in the united states. *PLoS medicine*, 3(9):e260, 2006.

National Archives. 1940 census. Available at https://1940census.archives.gov/, 2018.

NCHS. National health interview survey linked mortality files. Available at https://www.cdc.gov/nchs/data-linkage/mortality.htm, 2005.

Wayne B Nelson. *Applied life data analysis*, volume 577. John Wiley & Sons, 2005.

Fred Paccaud, C Sidoti Pinto, Alfio Marazzi, and Judith Mili. Age at death and rectangularisation of the survival curve: trends in switzerland, 1969-1994. *Journal of Epidemiology & Community Health*, 52(7): 412–415, 1998.

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

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

Martyn Plummer. Jags version 3.3.0 user manual. *International Agency for Research on Cancer, Lyon, France*, 2012.

Samuel Preston, Patrick Heuveline, and Michel Guillot. *Demography: measuring and modeling population processes*. Wiley-Blackwell, 2000.

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

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

Steven Ruggles, Catherine A Fitch, P Kelly Hall, and Matthew Sobek. Ipums-usa: Integrated public use microdata series for the united states. *Handbook of international historical microdata for population research. Minneapolis: Minnesota Population Center*, pages 259–284, 2000.

Argun Saatcioglu and John L Rury. 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, 2012.

Carl Schmertmann, Emilio Zagheni, Joshua R Goldstein, and Mikko Myrskylä. Bayesian forecasting of cohort fertility. *Journal of the American Statistical Association*, 109(506):500–513, 2014.

Paul D Sorlie, Eric Backlund, and Jacob B Keller. Us mortality by economic, demographic, and social characteristics: the national longitudinal mortality study. *American Journal of Public Health*, 85(7): 949–956, 1995.

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

Tzu Han Tai and Andrew Noymer. Models for estimating empirical gompertz mortality: With an application to evolution of the gompertzian slope. *Population Ecology*, pages 1–14, 2017.

Shripad Tuljapurkar and Ryan D Edwards. Variance in death and its implications for modeling and forecasting mortality. *Demographic Research*, 24:497, 2011.

James W Vaupel and Trifon I Missov. Unobserved population heterogeneity. *Demographic Research*, 2014.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432, 2017.

Kenneth W Wachter. *Essential Demographic Methods*. Harvard University Press, 2014.

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

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

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

John Wilmoth, Sarah Zureick, Vladimir Canudas-Romo, Mie Inoue, and Cheryl Sawyer. A flexible two-dimensional mortality model for use in indirect estimation. *Population Studies*, 66(1):1–28, 2012.

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

John R Wilmoth, Kirill Andreev, Dmitri Jdanov, Dana A Glei, C Boe, M Bubenheim, D Philipov, V Shkolnikov, and P Vachon. 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, 2007.

William E Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. 1990.