A General Age-Specific Mortality Model with An Example Indexed by Child or Both Child and Adult Mortality

Samuel J. Clark[1,2,*]

[1]Department of Sociology, The Ohio State University
[2]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand
[*]Contact: work@samclark.net, 206.303.9620

2018

Abstract

The majority of countries in Africa and nearly one third of all countries require mortality models to infer the complete age schedules of mortality that are required to conduct population estimates, projections/forecasts, and other tasks in demography and epidemiology. Models that relate child mortality to mortality at other ages are important because almost all countries have measures of child mortality. A general, parametrizable component model of mortality is defined using the singular value decomposition (SVD-Comp) and calibrated to the relationship between child or child/adult mortality and mortality at other ages in the observed mortality schedules of the Human Mortality Database. Cross validation is used to validate the model, and the predictive performance of the model is compared to that of the Log-Quad model, designed to do the same thing. Prediction and cross validation tests indicate that the child mortality-calibrated SVD-Comp is able to accurately represent the observed mortality schedules in the Human Mortality Database, is robust to the selection of mortality schedules used to calibrate it, and performs better than the Log-Quad Model. The child mortality-calibrated SVD-Comp can be used where child mortality is available but mortality at other ages is unknown.

1 Introduction

Complete age-specific mortality schedules are necessary inputs to a wide variety of formal demographic and epidemiological methods. A key example is the biennial World Population Prospects (WPP) (United Nations, Department of Economic and Social Affairs, Population Division, 2015b) produced by the UN Population Division. These are generally considered the reference population indicators and are widely used by other domestic and international agencies as inputs to estimation and modeling exercises. The WPP contains estimates of time-sex-age-specific mortality, fertility, and population size from 1950 to the present and forecasts of the same quantities to 2100 for all countries of the world. Consequently each WPP update must contain full age-specific mortality schedules covering the period 1950–2100.

Table 1: Countries or regions with no information on either child or adult mortality. UN countries and regions that do not have information on either child or adult mortality for the 2015 update of the World Population Prospects, with population and fraction of total population for which information is missing. *Reference:* United Nations, Department of Economic and Social Affairs, Population Division (2015c) tables I.1b (p 5) and I.1c (p 6).

	Child Mortality			Adult Mortality		
	Regions	Population (millions)	Percent Population	Regions	Population (millions)	Percent Population
World	1	1	0.0%	50	973	13.2%
Africa	1	1	0.0%	33	666	56.1%

Some countries in the developing world, particularly in Africa, do not yet have civil registration and vital statistic systems that function well enough to accurately report on either fertility or mortality. Focusing on mortality, Table 1 displays the number of countries or world regions for which there is no information on either child mortality or adult mortality, with Africa broken out. Because of the exhaustive coverage of household surveys investigating fertility and maternal/child health, essentially the whole world has at least some recent information on child mortality (Li, 2015). In contrast 50 countries around the world with a total population of nearly 1 billion people have no information on adult mortality, with the bulk of those in Africa – 33 countries with a total population of 666 million people.

Mortality models are used to solve this problem and produce full age schedules of mortality. Table 2 describes the number of countries or world regions for which the UN Population Division must use mortality models of some kind to produce either estimates of life expectancy at birth e_0 or full age schedules of mortality. Most African countries require mortality models for both, and globally 38.6% of countries require a model for e_0 and 32.6% for age-specific mortality.

The standard approach to generating complete age schedules of mortality for countries and areas with insufficient data is to take advantage of the fact that they do have information on child mortality. Typically, model life tables are used to extrapolate full mortality schedules from ${}_5q_0$ – this is what the UN Population Division does (making heavy use of the traditional Coale and Demeny (1966) model life tables), and the Institute for Health Metrics and Evaluation (IHME) uses variations on the Modifed Logit (Mod-Logit) model (Murray et al., 2003) to do the same.

Table 2: Countries and regions where mortality models are necessary to estimate life expectancy at birth (e_0) or age-specific mortality rates (ASMR). Counts of the number of UN countries and areas where mortality models were used to generate estimates of e_0 or age-specific mortality rates for the 2015 update of the World Population Prospects. *Reference:* United Nations, Department of Economic and Social Affairs, Population Division (2015a).

		e_0		ASMR	
	Countries/Regions	Count	Percent	Count	Percent
World	233	90	38.6%	76	32.6%
Africa	58	50	86.2%	50	86.2%

The commonly used model life table systems – Regional Model Life Tables and Stable Populations (Coale and Demeny, 1966), Life Tables for Developing Countries (United Nations, Department of Economic and Social Affairs, Population Division, 1982), Modified Logit Life Table System (Mod-Logit) (Murray et al., 2003; Wang et al., 2013) and Flexible Two Dimensional Mortality Model (Log-Quad) (Wilmoth et al., 2012) – combine a specific model structure and defined variable parameters with a set of fixed parameters that summarize the relationships between mortality at different ages in a set of observed life tables. All are *empirical* models in the sense that they summarize observed mortality and use that summary to produce predicted mortality schedules that are consistent with observed mortality. They come in both regional and continuous forms. The regional models identify and replicate commonly observed mortality patterns associated with geographic regions (and *de facto* time periods) and allow mortality to vary continuously within each region-specific pattern. In contrast the continuous models generate mortality patterns that vary smoothly. Both approaches are essentially two-parameter models. The regional models first identify a discrete region and then use effectively continuously varying life expectancy within each region to adjust the level of region-specific mortality. The continuous models have two continuously varying parameters, e.g. life expectancy, child mortality, or adult mortality.

Murray et al. (2003) enumerate three characteristics required of mortality models: 1) simplicity and ease of use, 2) comprehensive representation of the true variability in sex-age-specific mortality observed in real populations, and 3) validity that is well quantified by comparing age schedules of mortality predicted by the model to corresponding observed life tables. To those I would add: 1) generality with respect to the underlying model structure, 2) flexibility in terms of input parameters, and 3) an ability to handle a wide range of age groups, including very fine-grained, without having to fundamentally alter the structure of the model.

This work defines and describes a new SVD component-based mortality modeling framework that satisfies all of those requirements. The SVD-component framework provides a general, flexible way to model any demographic age schedule as a function of covariates or predictors that are related to age-specific variation in the age schedule. Here the SVD-component framework is demonstrated by creating a mortality model that predicts single-year of age mortality schedules using either ${}_5q_0$ or both $({}_5q_0, {}_{45}q_{15})$ as predictors, similar to both the Mod-Logit and Log-Quad models. The resulting model can be used to produce single year of age mortality schedules from ${}_5q_0$ alone that are consistent with observed mortality schedules, and this could be useful for those

like the UN Population Division who must manipulate full age schedules of mortality but only have observed values for ${}_5q_0$. The resulting SVD-component model performs better than the current state of the art two-parameter model (Log-Quad), provides predictions by single-year of age, and is easily extensible to include additional predictors beyond child and adult mortality.

The remainder of this article 1) reviews existing mortality models with an emphasis on those that use a dimension-reduction approach, 2) identifies and describes the empirical life tables used to develop the model, 3) develops and calibrates the model so that it reflects observed mortality across a wide range of settings and times, 4) uses a cross-validation approach to validate the model, 5) compares the performance of the model with that of the Log-Quad model, and 6) presents an example application to mortality data from two developing countries.

2 Mortality Models

Traditional model life tables (e.g. United Nations, Department of Economic and Social Affairs, Population Division, 1955; Ledermann, 1969; Coale and Demeny, 1966; United Nations, Department of Economic and Social Affairs, Population Division, 1982; Murray et al., 2003; Wilmoth et al., 2012; Wang et al., 2013) take an inductive, empirically-driven approach to identify and parsimoniously express the regularity of mortality with age based on observed relationships in large collections of high quality life tables. Some fertility models (e.g. Coale and Trussell, 1974; Lee, 1993) do the same. An alternative, sometimes deductive approach, can be found in the wide variety of parametric or functional-form mortality models (e.g. Gompertz, 1825; Makeham, 1860; Heligman and Pollard, 1980; Li and Anderson, 2009) that define age-specific measures of mortality in an analytical form, sometimes with interpretable parameters. Brass (1971) developed a new approach with his two-parameter 'relational' model that has been extended and refined in many ways, (for example Zaba, 1979; Murray et al., 2003). More recently the Log-Quad model of Wilmoth et al. (2012) combines empirical and functional-form approaches to mortality models.

Population forecasting has motivated another important family of related mortality models. Forecasting generates many iterations of both age-specific mortality and fertility into the future, and those are usually based on a summary of the corresponding age-specific mortality and fertility in the past. Hence there is an immediate need to represent full age schedules and their dynamics *compactly*. This led to the widespread use of dimension-reduction or data compression techniques to reduce the dimensionality of the problem so that only a few parameters are necessary to represent age schedules and their dynamics. Ledermann and Breas (1959) appear to have been the first to use principal components analysis (PCA) to summarize age-specific mortality and generate model life tables, and this approach was refined by many subsequent investigators, (e.g. Bourgeois-Pichat, 1962, 1990; Ledermann, 1969; United Nations, Department of Economic and Social Affairs, Population Division, 1982). Following the early use of PCA to build model life tables, PCA and related methods like the singular value decomposition (SVD) (e.g. Good, 1969; Stewart, 1993; Strang, 2009) have been widely used and refined by forecasters to create time series models of mortality and fertility (e.g. Bozik and Bell, 1987; Lee and Carter, 1992; Lee, 1993). Bell (1997) provides a comprehensive summary of this line of development in various fields, dominated by actuarial science and applications in forecasting.

The 'Lee-Carter' approach (Lee and Carter, 1992; Lee, 1993) has been widely used in demogra-

phy. The model as presented in Lee and Carter (1992) is

$$\ln(\mathbf{m}_{xt}) = \mathbf{a}_x + \mathbf{b}_x k_t + \epsilon_{xt} \ , \tag{1}$$

where x is age, t is time, **m** is a matrix of age, time-specific mortality rates, **a** is the time-constant vector of mean (over columns of **m**) logged age-specific mortality rates through time, and **b** is the time-constant first left singular vector from an SVD decomposition of the matrix of residuals generated by subtracting **a** from each column of **m**.

Fitting the model requires three separate steps: 1) calculate $\mathbf{a}_x$, 2) calculate the residuals $r_{xt} = \ln(\mathbf{m}_{xt}) - \mathbf{a}_x$, and 3) extract the first left singular vector from the SVD of **r** and calculate a value of k_t for each column of **m** that minimizes the elements ϵ_{xt} (k_t are essentially the elements of the first right singular vector multiplied by the first singular value of this SVD).

There are two conceptually separate elements to the Lee-Carter model, 1) a one-parameter (i.e. k_t) model of the full age-specific mortality or fertility schedule and 2) a time series model for that parameter. The temporal sequence of values taken by k_t is the focus of a stochastic time series model that is responsible for the temporal dynamics of the method, including the forecasts. Development of the time series model is previewed in earlier work by the authors (Carter and Lee, 1986).

Putting aside the time series model for k_t, it becomes clear that the structure of the Lee-Carter model appears to be a simplified version of the more complex age-period-cohort mortality model conceived earlier by Wilmoth and elaborated over a number of years (Wilmoth and Caselli, 1987; Wilmoth et al., 1989; Wilmoth, 1990)[1]. Wilmoth's model is designed to separate and identify age, period and cohort effects in an age $\times$ time matrix of mortality rates. The basic structure is $\log(m_x)$ = [mean model] + [residual model] with the final form

$$f_{ij} = \underbrace{\alpha_i + \beta_j}_{\text{mean model}} + \underbrace{\sum_{m=1}^{\rho} \phi_m \gamma_{im} \delta_{jm}}_{1^{\text{st}}\text{ residual model}} + \underbrace{\theta_k}_{2^{\text{nd}}\text{ residual model}} + \epsilon_{ij} \ , \tag{2}$$

where i is age, j is period, $k = (j - i)$ indexes cohorts, f is logged age-period-specific mortality $\log(m)$, α is an age effect, β is a period effect, the sum $\sum_{m=1}^{\rho} \phi_m \gamma_{im} \delta_{jm}$ is over a set of ρ rank-1 matrices from the SVD of the residuals remaining after the main effects are subtracted from f, and θ_k is a residual cohort effect remaining after subtracting both the main effects and the SVD approximation of the first residuals from f. This form first appears in Wilmoth et al. (1989).

The model is fit in three steps, effectively explaining ever more nuanced variation in a sequence of residuals: 1) calculate α_i and β_j such that they minimize the first residuals $r_{ij} = f_{ij} - (\alpha_i + \beta_j)$, 2) take the first ρ terms from the SVD of the matrix of residuals **r** and calculate the second residual $s_{ij} = r_{ij} - \sum_{m=1}^{\rho} \phi_m \gamma_{im} \delta_{jm}$, and 3) calculate values for the elements of θ_k such that they minimize $s_{ij} - \theta_k = \epsilon_{ij}$. The SVD or 'multiplicative' term $\sum_{m=1}^{\rho} \phi_m \gamma_{im} \delta_{jm}$ takes shape over several publications (Wilmoth and Caselli, 1987; Wilmoth et al., 1989; Wilmoth, 1990) to eventually be the standard SVD form that appears in the final model, with the first appearance of the SVD in Wilmoth et al. (1989).

[1] The core ideas underlying the Wilmoth model appear in his Ph.D. dissertation (Wilmoth, 1988), with further refinement in the following years, culminating in the English-language summary published in *Sociological Methodology* in 1990 (Wilmoth, 1990).

An examination of Equations 1 and 2 reveals the relationship between the Wilmoth and Lee-Carter models. To move from Wilmoth to Lee-Carter: 1) remove the main period effect β_j and the cohort effect θ_k and 2) take only the first term in the SVD approximation of the first residual. The SVD term then becomes $\phi_1\gamma_{i1}\delta_{j1}$, or dropping the $m = 1$ index, $\gamma_i(\phi\delta_j)$. Replacing Wilmoth's *i* and *j* with Lee-Carter's *x* and *t* and letting $k = \phi\delta$ makes the equivalence clear. In their 1992 publication Lee and Carter acknowledge that their model has much in common with the Wilmoth model. They cite Wilmoth by way of explaining the SVD 'solution' to calculating the elements of **b**, whereas this is just the simplest rank-1 form of the time-varying term in the model proposed by Wilmoth.

Motivated by the work of the UN Population Division that sometimes involves predicting full age schedules of mortality from child (and adult) mortality (Li, 2015), Wilmoth et al. (2012) present another adaptation of the original Wilmoth model, this time to generate model life tables as a function of ${}_5q_0$ or $({}_5q_0, {}_{45}q_{15})$. Adopting the nomenclature from log-linear models, this log-quadratic (Log-Quad) model has the form

$$\log(m_x) = a_x + b_x h + c_x h^2 + v_x k \ , \tag{3}$$

where *x* is age; *m* is age-specific mortality; *a*, *b*, and *c* are constant age-specific coefficients for the quadratic mean model, *h* is the input value of $\log({}_5q_0)$, *v* is an age-specific 'correction factor', and *k* is a coefficient for *v*. Correction factor values v_x are identified by calculating the SVD of the matrix of residuals that remain after the quadratic portion of the model is subtracted from life tables that are part of the Human Mortality Database (University of California, Berkeley and Max Planck Institute for Demographic Research) and using the resulting first left singular vector as a starting point[2]. Thus, the Log-Quad model has the now familiar mean/residual form of the original Wilmoth model and the structure of the residual model is a one-term version of the SVD form originally proposed by Wilmoth et al. (1989). The Log-Quad's contribution is an innovative new mean model that takes advantage of the empirically observed curvilinear relationship between child mortality and mortality at other ages. The Log-Quad model is elegant, simple, and parsimonious – one $({}_5q_0)$ or two ($_5q_0$ and *k*)[3] parameters – and it performs well, accurately representing a wide range of life tables, including life tables with very low mortality, and generally outperforming all other existing model life tables (Wilmoth et al., 2012).

Recently other investigators have worked on a variety of matrix-summary approaches to characterize the variability in mortality rates, but none of their work has been as widely used as the Lee-Carter model. Working independently, Fosdick and Hoff (2012) develop an explicitly statistical 'separable factor analysis' model to summarize mortality in the HMD, and at its core this is similar to the SVD term in Wilmoth's model. Also working independently, I developed a 'component model' of mortality inspired by the use of matrix factorization methods and the fast Fourier transform in image compression (Clark, 2001). The component model is a simple linear sum of independent, age-varying vectors (components) that when combined with appropriate weights can closely approximate age-specific mortality schedules. This model has the simple basic form

$$\mathbf{m} = \sum_{i=1}^{c} w_i \mathbf{u}_i + \mathbf{r} \ , \tag{4}$$

[2]The first left singular vector of the HMD residuals are massaged slightly to ensure all elements of *v* are positive and 'smooth'.

[3]If desired, *k* is chosen so that the resulting mortality schedule matches an input value ${}_{45}q_{15}$.

where $\mathbf{m}$ is a vector of age-specific mortality rates, $\mathbf{u}_i$ are a set of c vectors containing age-varying values identified in a set of observed mortality rates, w_i are weights, and $\mathbf{r}$ is a vector of residuals. This is similar to Ledermann's original use of factor analysis to build a system of model life tables based on factors resulting from a PCA decomposition of a matrix of age-specific mortality rates (Ledermann and Breas, 1959; Ledermann, 1969) and the PCA-based model underlying the UN model life tables (United Nations, Department of Economic and Social Affairs, Population Division, 1982) – both of which have the mean/residual structure of the Wilmoth models because they use PCA operating on a *centered* data cloud. The component model has been used to summarize mortality data from the INDEPTH Network using PCA-derived components (Clark, 2001; INDEPTH Network, 2002; Clark et al., 2009), similarly for the HMD (Clark and Sharrow, 2011b,a), and more recently in work on small-area estimates of mortality (Alexander et al., 2016). This approach combines a simple linear model with PCA, SVD or similar methods to concentrate information along a few dimensions, see (Clark, 2015) for a detailed discussion.

The component model is similar to the SVD-inspired '1$^{\text{st}}$ residual model' term in Wilmoth's Equation 2. However, neither Wilmoth nor subsequent investigators identify or develop the relationship between the SVD decomposition of a matrix of mortality rates and the column-wise, weighted-sum model in Equation 4. A key conceptual difference between the two approaches is that Equation 4 does not have a 'mean model', and consequently the factors identified by the SVD model everything, not just the residual as in all of the Wilmoth-inspired models. The first component $\mathbf{u}_1$ is effectively the mean age-specific mortality schedule and its weight reflects the overall level of mortality. The remaining components $\mathbf{u}_i$ for $i > 1$ define deviations from the average age pattern, independent of level. All of this follows directly from the properties of the SVD and a substantive interpretation of both the left and right singular vectors when applied to demographic age schedules (Clark, 2015). Additionally, the weights are viewed as continuously varying parameters that can be the object or output of additional models - e.g. clustered using objective clustering methods to identify groups of similar age schedules, estimation using either traditional or Bayesian methods, or predicted from covariates that vary systematically with age schedules, as I demonstrate below.

Finally, we recently applied the component model to HIV-related mortality in countries with large HIV epidemics (Sharrow et al., 2014). In that article we demonstrate that the weights in Equation 4 vary systematically with HIV prevalence. We took advantage of that fact to build a model that predicts three weights as a function of HIV prevalence and then predicts mortality age schedules from the predicted weights using Equation 4. The resulting 'HIV-calibrated' component model uses the weights as a link between HIV prevalence and full age schedules of mortality.

Below I describe how the singular value decomposition can be used to develop a general modeling framework for demographic age schedules. This framework has the important advantages of being 1) straightforward and easy to understand and use, 2) general and applicable to any demographic age schedule, 3) able to incorporate covariates or predictors in a unified way, and 4) able to handle age groups of any granularity, e.g. one-year or five-years, in the same way. This framework is demonstrated by creating and validating an accurate one or two-parameter mortality model based on age-patterns of mortality contained in the Human Mortality Database (University of California, Berkeley and Max Planck Institute for Demographic Research).

3 Data

3.1 Human Mortality Database Life Tables - HMD

The Human Mortality Database (HMD) (University of California, Berkeley and Max Planck Institute for Demographic Research) contains rigorously cleaned, checked and validated information on deaths and exposure from a number of mainly developed countries where "death registration and census data are virtually complete". The data are aggregated and presented in a wide variety of formats. The objective of this analysis is to capture and characterize as much variability in age-specific mortality as possible, and consequently I chose to use the 1×1 HMD life tables for each sex. Those provide all columns of a standard life table for single calendar years by single year of age from $0 \rightarrow 110+$. Each country provides data for different historical periods, and some countries are subdivided into more specific subpopulations. In the latter situation a 'national population' life table is typically provided that aggregates across the subgroups. Both the national and subgroup populations are included in this analysis to maximize the variability in age-specific mortality schedules in the overall dataset. A few of the 1×1 life tables from the HMD contain problems: 1) the life tables for Belgium 1914-1918 for both sexes contain no data, and 2) the female life tables for Iceland (ISL) 1852, and New Zealand Mauri (NZL_MA) 1949, 1956, and 1959 display implausible mortality at older ages. All of those life tables are excluded. Table 3 contains an organized list of the life tables included in this analysis. There are 4,610 life tables for each sex, 9,220 in total. The HMD data used in this analysis are contained in the file at `http://www.mortality.org/hmd/zip/all_hmd/hmd_statistics.zip` downloaded on Local cached HMD.

3.2 Model Scales

This analysis is conducted on life table probabilities of dying for those who survive to the beginning of each one-year age group. Single year probabilities ${}_1q_x$ are taken directly from the HMD life tables, five-year probabilities ${}_5q_x$ are calculated as ${}_5q_x = 1 - \prod_{a=x}^{x+4}(1 - {}_1q_a)$, and ${}_{45}q_{15}$ is calculated as ${}_{45}q_{15} = 1 - \prod_{a=15}^{59}(1 - {}_1q_a)$. 'Child mortality' refers to ${}_5q_0$ and 'adult mortality' refers to ${}_{45}q_{15}$.

The natural scale of the models described below is the full real line, so life table probabilities of dying q are transformed using the *logit* function $\text{logit}(x) = \ln\left(\frac{x}{1-x}\right)$ so that their transformed values occupy the full real line. Outputs from the models are transformed back to the probability scale with range [0,1] using the *expit* function $\text{expit}(x) = \frac{e^x}{1+e^x}$, inverse of the *logit*.

4 Methods

This section is divided into four topics: 1) the general SVD-component modeling framework is described based on attributes of the SVD, 2) a specific application of the SVD-component modeling framework is developed to create the *SVD-Comp* mortality model that relates child or child and adult mortality to mortality at all other ages using data from the Human Mortality Database (University of California, Berkeley and Max Planck Institute for Demographic Research), 3) a testing and validation strategy is developed to test SVD-Comp and compare its performance to the Log-Quad model, and 4) an example application of SVD-Comp to mortality data not included in the HMD is described.

Table 3: Life Tables. 4,610 consistent 1 $\times$ 1 (single-year in both calendar and age) life tables downloaded from the Human Mortality Database on Local cached HMD.

Country/Population	Abbreviation	Years Covered	Total Life Tables
Australia	AUS	1921–2014	94
Austria	AUT	1947–2017	71
Belgium	BEL	1841–1913	73
Belgium	BEL	1919–2015	97
Bulgaria	BGR	1947–2010	64
Belarus	BLR	1959–2016	58
Canada	CAN	1921–2011	91
Switzerland	CHE	1876–2016	141
Chile	CHL	1992–2008	17
Czechia	CZE	1950–2016	67
East Germany	DEUTE	1956–2015	60
Germany	DEUTNP	1990–2015	26
West Germany	DEUTW	1956–2015	60
Denmark	DNK	1835–2016	182
Spain	ESP	1908–2016	109
Estonia	EST	1959–2017	59
Finland	FIN	1878–2015	138
France – Civilian Population	FRACNP	1816–2016	201
France – Total Population	FRATNP	1816–2016	201
England and Wales – Civilian National Population	GBRCENW	1841–2016	176
England and Wales – Total Population	GBRTENW	1841–2016	176
Northern Ireland	GBR_NIR	1922–2016	95
United Kingdom	GBR_NP	1922–2016	95
Scotland	GBR_SCO	1855–2016	162
Greece	GRC	1981–2013	33
Croatia	HRV	2002–2016	15
Hungary	HUN	1950–2017	68
Ireland	IRL	1950–2014	65
Iceland	ISL	1838–1851	14
Iceland	ISL	1853–2016	164
Israel	ISR	1983–2014	32
Italy	ITA	1872–2014	143
Japan	JPN	1947–2016	70
Korea	KOR	2003–2016	14
Lithuania	LTU	1959–2017	59
Luxembourg	LUX	1960–2014	55
Latvia	LVA	1959–2017	59
Netherlands	NLD	1850–2016	167
Norway	NOR	1846–2014	169
New Zealand – Maori	NZL_MA	1948–1948	1
New Zealand – Maori	NZL_MA	1950–1955	6
New Zealand – Maori	NZL_MA	1957–1958	2
New Zealand – Maori	NZL_MA	1960–2008	49
New Zealand – Non-Maori	NZL_NM	1901–2008	108
New Zealand	NZL_NP	1948–2013	66
Poland	POL	1958–2016	59
Portugal	PRT	1940–2015	76
Russia	RUS	1959–2014	56
Slovakia	SVK	1950–2014	65
Slovenia	SVN	1983–2014	32
Sweden	SWE	1751–2016	266
Taiwan	TWN	1970–2014	45
Ukraine	UKR	1959–2013	55
The United States of America	USA	1933–2016	84

4.1 Relevant Characteristics of the Singular Value Decomposition

This section summarizes from Clark (2015). The SVD (e.g. Good, 1969; Stewart, 1993; Strang, 2009) is a matrix factorization method that decomposes a matrix $\mathbf{X}$ into three matrix factors with special properties:

$$\mathbf{X} = \mathbf{USV}^{\mathsf{T}} \ . \tag{5}$$

$\mathbf{U}$ is a matrix of 'left singular vectors' (LSVs) arranged in columns, $\mathbf{V}$ is a matrix of 'right singular vectors' (RSVs) arranged in columns, and $\mathbf{S}$ is a diagonal matrix of 'singular values' (SVs). The LSVs and RSVs are independent and have unit length. If one views the columns of $\mathbf{X}$ as a set of dimensions, then the rows of $\mathbf{X}$ locate points defined along those dimensions – the data cloud. The RSVs define a new set of dimensions that line up with the axes of most variation in the data cloud. The first RSV points from the origin to the data cloud, or if the cloud is around the origin, then it points along the line of maximum variation within the cloud. The remaining RSVs are orthogonal to the first and each other and line up with successively less variable dimensions within the cloud. The elements of the LSVs are values that correspond to the projection of each point along the new dimensions defined by the RSVs. The SVs effectively stretch the new dimensions defined by the RSVs in accordance with the variation in the cloud along each RSV. The numeric value of each SV is the square root of the sum of squared distances from the origin to each point along the corresponding SVD dimension, and their squares sum to the total sum of squared distances from the origin to each point along all of the original dimensions.

The basic form of the SVD in Equation 5 can be rearranged to yield two new useful expressions

$$\mathbf{X} = \sum_{i=1}^{\rho} s_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}} \tag{6}$$

and

$$\mathbf{x}_\ell = \sum_{i=1}^{\rho} s_i v_{\ell i} \mathbf{u}_i \ , \tag{7}$$

where $\mathbf{u}_i$ are LSVs, $\mathbf{v}_i$ are RSVs, s_i are SVs, ρ is the rank of $\mathbf{X}$, $\mathbf{x}_\ell$ are columns of $\mathbf{X}$, and $v_{\ell i}$ are the elements of RSV $\mathbf{v}_i$, see Appendix A (on-line appendices). Equation 6 says that $\mathbf{X}$ can be written as a sum of rank-1 matrices, each created from one of the LSVs by applying weights in the form of the elements of the corresponding RSV. Equivalently, Equation 7 says that each column $\mathbf{x}_\ell$ of $\mathbf{X}$ can be written as the weighted sum of the LSVs with the weight for each being the ℓ^{th} element of the corresponding RSV[4]. The LSVs and SVs are constant, so the the weights are the 'variables' in these expressions, and their values determine how much of each LSV is added to the mixture to represent the original data. Finally because the LSVs are independent, OLS regression can be used to estimate models that relate $\mathbf{x}_\ell$ to the LSVs. If the constant is constrained to be zero, then the coefficients are equal to $s_i v_{\ell i}$.

Because the RSVs define successively less variable dimensions in the data cloud, the first term in Equations 6 and 7 contains the most information and subsequent terms contain less and less (Golub et al., 1987). Including all ρ terms replicates the original data matrix $\mathbf{X}$ or any of its columns $\mathbf{x}_\ell$ exactly, while including only the first few terms provides a good approximation.

4.2 SVD Component Model – 'SVD-Comp'

Given an $A \times L$ matrix $\mathbf{Q}$ of mortality schedules for each sex, calculate the SVD($\mathbf{Q}_z$) = $\mathbf{U}_z\mathbf{S}_z\mathbf{V}_z^{\mathsf{T}}$. Using the resulting factors as in Equation 7, each A-element mortality schedule $\mathbf{q}_{z\ell}$ is approximated

[4] This is the expression used to model the first residual in Wilmoth's age/period/cohort model, Equation 2.

as the c-term sum

$$\mathbf{q}_{z\ell} \approx \sum_{i=1}^{c} v_{z\ell i} \cdot s_{zi}\mathbf{u}_{zi} \ , \tag{8}$$

where A is the number of age groups and rows in $\mathbf{Q}_z$; L is the number of life tables and columns in $\mathbf{Q}_z$; $z \in \{\text{female}, \text{male}\}$; $c \leq \rho$, the rank of $\mathbf{Q}_z$; and $\ell \in \{1 \ldots L\}$ indexes mortality schedules (Golub et al., 1987). The A-element LSVs $\mathbf{u}_{zi}$ and the SVs s_{zi} are constant across all mortality schedules. Because $c \leq \rho$, the sum on the right is an approximation of the mortality schedule, hence the '$\approx$'. As is clear just below in Section 4.4, $c = 4$ is sufficient to make the approximation almost perfect across the entire HMD. Viewed as a data compression technique, all 4,610 sex-specific mortality schedules in the HMD can be very closely approximated with just four age-varying components – a greater than 99.9% reduction in the volume of data required to represent the HMD. The elements that vary among mortality schedules are the RSVs $\mathbf{v}_{zi}$ whose elements $v_{z\ell i}$ are the weights in the sum. This is a continuously varying model like Mod-Logit (Murray et al., 2003) and Log-Quad (Wilmoth et al., 2012) rather than a regional model like the Coale & Demeny (Coale and Demeny, 1966) and UN (United Nations, Department of Economic and Social Affairs, Population Division, 1982) model life tables.

Figure 2, below in Section 5, displays the scaled left singular vectors $s_{zi}\mathbf{u}_{zi}$ obtained from the SVD of the matrix of logit-scale ${}_1q_x$ values contained in the Human Mortality Database (University of California, Berkeley and Max Planck Institute for Demographic Research). The SVD-Comp model is simply a weighted sum of those components. The first component represents the average shape and scale of human mortality by age and the remaining three add age-specific modifications to that basic shape – i.e. all values of the first component are negative (because of the logit transformation), while components two through four all cross the x-axis.

When the $v_{z\ell i}$ are replaced by values that can be related to covariates, as they are just below, this becomes a highly flexible modeling framework that can be used inductively like traditional model life tables to produce a mortality model that generates age schedules of mortality that are consistent with a collection of observed mortality schedules, or it can be used deductively to generate new age schedules based on a theoretical understanding of how a covariate should affect each component in the model. In general, the age pattern of the scaled LSVs in the sum can be interpreted and manipulated theoretically, see Figure 2 and the results in Section 5.2.

4.3 Parametrization using ${}_5\mathbf{q}_0$ and $({}_5\mathbf{q}_0, {}_{45}\mathbf{q}_{15})$

Equation 8 describes a relationship between the elements of the RSVs and the age schedule of mortality. Consequently, if a covariate is related to the age schedule of mortality, it will necessarily also have a relationship with the elements of the RSVs, particularly the first few RSVs corresponding to the SVD-defined dimensions that capture the majority of the variability in the data cloud formed by the HMD life tables. It is possible to take advantage of this fact to define and estimate models that relate the elements of the RSVs to child mortality and adult mortality. These take the form

$$v_{z\ell i} = f_{zi}({}_5q_{0\,z\ell}) \tag{9}$$

and

$$v_{z\ell i} = f_{zi}({}_5q_{0\,z\ell},\ {}_{45}q_{15\,z\ell}) \ , \tag{10}$$

where, again, $z \in \{\text{female}, \text{male}\}$, $i \leq \rho$ indexes the RSVs, and $\ell \in \{1 \ldots L\}$ indexes both the elements of the RSVs and the values of child and adult mortality, one for each sex-specific mortality

schedule. There is a separate model f_{zi} for each sex-specific RSV, and these models can be used to produce predicted values for the weights in Equation 8 using new values for ${}_5q_{0\,z}$ and ${}_{45}q_{15\,z}$.

Following our earlier work (Sharrow et al., 2014; INDEPTH Network, 2002), the final model for any age schedule of mortality probabilities $\mathbf{q}_z$ associated with given values for a set of weights $\widehat{w}_{zi} = f_{zi}({}_5q_{0\,z})$ or $\widehat{w}_{zi} = f_{zi}({}_5q_{0\,z},\ {}_{45}q_{15\,z})$ is

$$\widehat{\mathbf{q}}_z = \sum_{i=1}^{c} \widehat{w}_{zi} \cdot s_{zi}\mathbf{u}_{zi}\ . \tag{11}$$

Equation 11 relates either child mortality ${}_5q_0$ or both child and adult mortality $({}_5q_0,\ {}_{45}q_{15})$ to full age schedules of mortality according to the patterns of those relationship that exist in the original set of HMD life tables $\mathbf{Q}$ using a very compact approximation.

This is a fully general approach to predicting mortality or any other demographic age schedules. Equations 9 and 10 can be replaced with models that summarize the relationships between any covariate and elements of the RSVs and weights, and age can be aggregated into any age groups – that simply requires recalculating the SVD on the age-aggregated data set.

4.4 Calibrating SVD-Comp to the Relationship between ${}_5\mathbf{q}_0$ and Mortality at Other Ages in the HMD

All computation is carried out using the R statistical programming environment (R Core Team, 2016; R Foundation for Statistical Computing, 2016).

4.4.1 Calibration SVDs.

The life tables of the HMD are arranged into two $A \times L$ matrices $\mathbf{Q}_z$ of single-year, age-specific life table probabilities of dying ${}_1q_x$, one for each sex. A = number of age groups = 110, L = number of life tables = 4,610, and $z \in \{\text{female, male}\}$. The SVD[5] of each $\mathbf{Q}_z$ yields ρ LSVs $\mathbf{u}_{zi}$, RSVs $\mathbf{v}_{zi}$, and SVs $\mathbf{s}_z$. To ensure that all age groups have approximately the same influence when calculating the SVDs, each mortality schedule is offset from the origin[6] by -10, and the offset is added back to predicted mortality schedules. Four of the new dimensions identified by each SVD are retained, i.e. $c = 4$ in Equation 11. For females those account for 0.998328, 0.000936, 0.000071, and 0.000058 of the total sum of squares, respectively, or together 0.999392. For males, 0.998595, 0.000824, 0.000103, and 0.000052 and together 0.999575. Appendix C (online appendices) contains additional information on the total sum of squares explained by each component of the SVD.

4.4.2 Models for Predicting Weights.

Based on Equations 9 and 10, regression models are defined that relate the RSVs $\mathbf{v}_{zi}$ to ${}_5q_{0\,z}$ and ${}_{45}q_{15\,z}$. Scatterplots of the elements of the RSVs versus logit$({}_5q_0)$ in Figures E.1 and E.2

[5] SVDs calculated using the svd function in the base package of R.

[6] This ensures that the whole data cloud is separated from the origin by an amount that is substantially greater than the typical value of each logit-transformed mortality rate, and therefore each age group has roughly equivalent leverage in the optimization required to identify the first new dimension of the SVD. The remaining dimensions are effectively identified on a centered data cloud.

in Appendix E (on-line appendices) make it clear that the relationships are not linear or simple. With no theory to guide the choice of predictors, I tried all combinations of simple transformations of $\text{logit}({}_5q_0)$ and $\text{logit}({}_{45}q_{15})$ and their interactions. The resulting models explain almost all the variance in the elements of $\mathbf{v}_1$ ($R^2 \approx 97\%$ for both sexes), the vast majority of the variance in the elements of $\mathbf{v}_2$ ($R^2 \approx 87\%$ for both sexes), and between one third and one half of the variance in the elements of $\mathbf{v}_3$ an d $\mathbf{v}_4$. Additionally, I tried to avoid overfitting or creating odd boundary effects in the predicted values that would have made out-of-sample predictions immediately implausible. These models behave sensibly up to the edges of the sample. The final models are

$$\begin{aligned} v_{z\ell i} = c_{zi} &+ \beta_{z1i} \cdot {}_5q_{0\,z\ell} + \beta_{z2i} \cdot \text{logit}({}_5q_0)_{\,z\ell} + \beta_{z3i} \cdot \text{logit}({}_5q_0)^2_{z\ell} + \beta_{z4i} \cdot \text{logit}({}_5q_0)^3_{z\ell} \\ &+ \beta_{z5i} \cdot {}_{45}q_{15\,z\ell} + \beta_{z6i} \cdot \text{logit}({}_{45}q_{15})^2_{z\ell} + \beta_{z7i} \cdot \text{logit}({}_{45}q_{15})^3_{z\ell} \\ &+ \beta_{z8i} \cdot [\text{logit}({}_5q_0)_{\,z\ell} \times \text{logit}({}_{45}q_{15})_{\,z\ell}] + \epsilon_{z\ell i}\,, \end{aligned} \tag{12}$$

where $i \in \{1:4\}$ indexes the SVD dimensions and ℓ indexes mortality schedules and elements of $\mathbf{v}_{zi}$. OLS regression is used to estimate coefficients for the eight regression models defined in Equation 12, and the estimated values are contained in Appendix D (on-line appendices) Tables D.1 and D.2. Using new values for both ${}_5q_0$ and ${}_{45}q_{15}$ as inputs, these models are used to predict values for the weights in Equation 11 – i.e. for prediction replace $v_{z\ell i}$ on the left hand side with $\widehat{w}_{zi}$.

4.4.3 Models for Adult Mortality.

To accommodate a one-parameter model that uses only ${}_5q_0$ as an input, a regression model is defined that relates adult mortality $\text{logit}({}_{45}q_{15})_z$ to child mortality ${}_5q_{0\,z}$. The scatterplot of $\text{logit}({}_{45}q_{15})$ versus $\text{logit}({}_5q_0)$ in Figure E.3 in Appendix E (on-line appendices) reveals a slightly complicated relationship that is neither linear nor systematically curvilinear. Again without theory as a guide, I tried a variety of models including various simple transformations of ${}_5q_0$. The resulting models explain most of the variance in $\text{logit}({}_{45}q_{15})$ (R^2 = 93% for females and 79% for males). The final models are

$$\begin{aligned} \text{logit}({}_{45}q_{15})_{\,z\ell} = c_z &+ \beta_{z1} \cdot {}_5q_{0\,z\ell} + \beta_{z2} \cdot \text{logit}({}_5q_0)_{\,z\ell} \\ &+ \beta_{z3} \cdot \text{logit}({}_5q_0)^2_{z\ell} + \beta_{z4} \cdot \text{logit}({}_5q_0)^3_{z\ell} + \epsilon_{z\ell}\,. \end{aligned} \tag{13}$$

OLS regression is used to estimate coefficients for the two regression models defined by Equation 13, and the estimated coefficients are contained in Appendix D (on-line appendices) Table D.3. This model is used to predict values for ${}_{45}q_{15}$ when only ${}_5q_0$ is supplied as an input. Then both the input value for ${}_5q_0$ and the predicted value for ${}_{45}q_{15}$ are used in Equation 12 to predict the weights in Equation 11.

4.4.4 Models for Mortality in the First Year of Life.

Figure E.4 displays the relationship between $\text{logit}({}_1q_0)$ and $\text{logit}({}_5q_0)$. Mortality falls very rapidly in the first few years of life. Using the child mortality rate ${}_5q_0$, a five-year summary of mortality between ages 0 and 5, as a predictor of single-year mortality within that same five-year age group is relatively uninformative. Experimentation reveals that ${}_5q_0$ predicts ${}_1q_1$ through ${}_1q_4$ well and ${}_1q_0$ slightly less well. The prediction of ${}_1q_0$ can be improved by modeling the relationship between

$\text{logit}({}_1q_0)$ and $\text{logit}({}_5q_0)$ separately as

$$\text{logit}({}_1q_0)_{z\ell} = c_z + \beta_{z1} \cdot \text{logit}({}_5q_0)_{z\ell} + \beta_{z2} \cdot \text{logit}({}_5q_0)^2_{z\ell} + \epsilon_{z\ell} \; . \tag{14}$$

OLS regression is used to estimate the coefficients of this model, displayed in Appendix D (on-line appendices) Table D.4. The model explains essentially all the variance in $\text{logit}({}_1q_0)$ ($R^2 > 99\%$ for both sexes) and is used to predict values for ${}_1q_0$ directly from the input value of ${}_5q_0$.

4.5 Using the Model

The full model is used in the following way:

1. Identify input values for ${}_5q_0$ and optionally ${}_{45}q_{15}$ and transform them to the logit scale. If ${}_{45}q_{15}$ is not available, predict $\text{logit}({}_{45}q_{15})$ using the input value for ${}_5q_0$ and the regression coefficients corresponding to Equation 13.
2. Use the input values for $\text{logit}({}_5q_0)$ and $\text{logit}({}_{45}q_{15})$ obtained in step 1 and the regression coefficients estimated using Equation 12 to predict values for the weights $\widehat{w}_{zi}$ defined Equation 11.
3. Insert the weights predicted in step 2 into Equation 11 to calculate a predicted age schedule of mortality probabilities $\widehat{\mathbf{q}}$ on the logit scale.
4. If desired, improve the prediction of $\text{logit}({}_1q_0)$ using the regression coefficients corresponding to Equation 14 to directly predict $\text{logit}({}_1q_0)$ from the input value of $\text{logit}{}_5q_0$ from step 1. Replace the first element of $\widehat{\mathbf{q}}$ with this predicted value for $\text{logit}({}_1q_0)$.
5. Add 10 to each element of $\widehat{\mathbf{q}}$ to account for the offset used when calculating the SVDs of the HMD mortality schedules.
6. Take the expit of $\widehat{\mathbf{q}}$ to yield single-year age-specific probabilities of dying on the probability scale.

4.6 Model Validation

The general sensitivity of the model to exactly which mortality schedules are used for calibration is assessed using a cross validation approach. Fifty random samples of 50% of the HMD mortality schedules are drawn, the model is calibrated with each using the calibration process described just above in Section 4.4, and all of the HMD mortality schedules are predicted. For each of the 50 models, prediction errors are calculated for all mortality schedule as the difference $\mathbf{q}_\ell - \widehat{\mathbf{q}}_\ell$. The error distributions of the in-sample and out-of-sample mortality schedules are summarized and compared.

In order to investigate how sensitive the overall modeling approach is to the number of mortality schedules used to calibrate the model, another cross validation exercise is conducted with varying sample sizes. For each sample fraction from 10% to 90% in 20% increments, 50 random samples are drawn from the HMD life tables . As above, the model is calibrated using each sample and all of the HMD mortality schedules are predicted, errors calculated, and error distributions for in- and out-of-sample mortality schedules are summarized and compared.

4.7 Comparing Performance of SVD-Comp and the Log-Quad Model

The Log-Quad model (Wilmoth et al., 2012) is the state-of-the-art mortality model relating child and adult mortality to full age schedules of mortality. I compare prediction errors produced by both the Log-Quad and SVD-Comp models. I use the Log-Quad model as published and the R code provided by Wilmoth et al. (2012) to produce predicted ${}_5q_x$ values for each of the HMD mortality schedules using either ${}_5q_0$ or both ${}_5q_0$ and ${}_{45}q_{15}$ as inputs. The Log-Quad model predicts mortality in five-year age groups. To accommodate the one-year age groups (${}_1q_x$) predicted by the SVD-Comp model, I use standard life table methods to transform predicted single-year to five-year ${}_5q_x$ values. I summarize the distribution of errors $\mathbf{q}_\ell - \widehat{\mathbf{q}}_\ell$ produced by both models in various ways. Comparisons are made only for predictions using the same inputs for both models, either ${}_5q_0$ alone or both $({}_5q_0, {}_{45}q_{15})$.

I also summarize the overall error produced by each model across all of the mortality schedules in the HMD. This is done by taking the absolute value of each year-sex-age-specific error and then summing the resulting absolute errors across all ages and years for each sex. This produces a single number – the total absolute error – that indicates the overall difference between the predicted and actual values for all years and ages. In addition to this I present total absolute errors in e_0.

To assess age-specific errors in $\widehat{\mathbf{q}}$ and life table quantities derived from $\widehat{\mathbf{q}}$, $\widehat{\mathbf{q}}_\ell$ are predicted with both SVD-Comp and Log-Quad using ${}_5q_0$ from each HMD life table as input. Full life tables are constructed from $\widehat{\mathbf{q}}_\ell$ and these are compared to the life tables in the HMD[7]. Age-specific weights are constructed from the l_x columns of the HMD life tables by summing l_x across all HMD life tables in five-year age intervals and then dividing each age-specific sum by the total across all ages. The resulting weights correspond to the proportionate l_x age structure of the HMD life tables. Weighted age-specific absolute errors in $\widehat{\mathbf{q}}$ and $\widehat{\mathbf{e}}$ are calculated by summing absolute errors in ${}_5\widehat{q}_x$ and $\widehat{e}_x$ at five year age intervals across all life tables in the HMD and then multiplying by the corresponding age-specific weight. The weighted age-specific errors in ${}_5\widehat{q}_x$ are a refinement on the overall errors in ${}_5\widehat{q}_x$, described just above, and reveal how close each model comes to replicating ${}_5q_x$ at each age. The weighted age-specific errors in $\widehat{e}_x$ provide an age-specific summary of the errors at each age in the derived life table columns that are necessary to calculate e_x - i.e. all of them.

4.8 Application to Mexico and South Africa

SVD-Comp and Log-Quad are used to predict age-specific mortality rates for Mexico 1983–1985 and South Africa 2005 using both child and adult mortality as inputs. Data for Mexico come from the Human Life Table Database (Max Planck Institute for Demographic Research et al., Downloaded August 2018), and data for South Africa from the World Health Organization's (WHO) Global Health Observatory data repository (World Health Organization, Downloaded August 21, 2018).

Mexico was chosen because it is a developing country with reasonable data and generally low but otherwise unremarkable mortality. South Africa was chosen because it is a developing country with a unique age-specific mortality schedule during the late 1990s and early 2000s. HIV/AIDS

[7] The SVD-Comp life tables are constructed using standard procedures in one-year age groups with ${}_na_x$ values taken from the HMD life tables. The Log-Quad life tables are constructed using R code provided by Wilmoth et al. (2012) in five-year age groups.

caused many deaths at very young and adult ages giving rise to a characteristic bulge in mortality at adult ages. Because both Log-Quad and SVD-Comp are calibrated using the HMD which does not contain life tables with HIV/AIDS-related mortality, both models are expected to do reasonably well with Mexico, but neither is expected to follow the HIV/AIDS-related mortality bulge in South Africa.

5 Results

5.1 Data and Fits

To provide a sense of the mortality data contained in the HMD and the fits produced by the SVD-Comp model, Figure 1 displays ${}_1q_x$ on the logit scale for Sweden in 1751 and Austria in 1990, with both data and predicted values produced by SVD-Comp using ${}_5q_0$ alone as an input.

5.2 Factors of the SVD

Figure 2 and Table B.1 in Appendix B (on-line appendices) present the sex-specific LSVs from the SVD of the full set of HMD mortality schedules scaled by their corresponding singular values, $s_i\mathbf{u}_i$ (ignoring the index for sex z). All elements of $s_1\mathbf{u}_1$ are negative so that $s_1\mathbf{u}_1$ captures the underlying 'average' shape of the mortality profile with age. Weights applied to $s_1\mathbf{u}_1$ move this underlying mortality profile up and down and hence control the overall level of mortality. The remaining $s_i\mathbf{u}_i$ all cross the x-axis and therefore represent age-specific deviations from the overall underlying pattern. These scaled left singular vectors are the components used in the weighted sum in Equation 11. Figure 2 also displays smoothed[8] versions of the scaled LSVs. One can use the smoothed versions to make the predicted mortality schedules smoother.

5.3 Calibration Relationships

Figures E.1 through E.4 in Appendix E (on-line appendices) display the data and predicted values from the models in Equations 12, 13, and 14. The corresponding estimated coefficients based on the whole HMD and used to calculate the predicted values in the figures are contained in Tables D.1, D.2, D.3, and D.4 in Appendix D (on-line appendices). Figures E.1 and E.2 in Appendix E (on-line appendices) contain scatterplots of the RSV element values versus logit(${}_5q_0$). The figures display both data and values predicted from Equation 12 using logit(${}_5q_0$) and logit(${}_{45}q_{15}$) predicted from the model in Equation 13 as inputs. There are clear, quasilinear relationships between the elements of the RSVs and logit(${}_5q_0$). Figure E.3 in Appendix E (on-line appendices) displays logit(${}_{45}q_{15}$) versus logit(${}_5q_0$), along with the predicted values from Equation 13. Finally, Figure E.4 in Appendix E (on-line appendices) displays logit(${}_1q_0$) versus logit(${}_5q_0$), along with predicted values from Equation 14.

5.4 Cross Validation Prediction Errors

Figure 3 displays sex-age-specific boxplots of the error distribution for one-year age groups from the first cross validation using 50 samples of 50% of the HMD to calibrate the SVD-Comp model. The errors are generally very small and centered around zero through roughly age 60. At older ages the size of the errors increases, and the median drifts slightly away from zero in a positive

[8] For components $i \in \{2, 3, 4\}$, kernel smoother with Gaussian kernel and bandwidth = $i + 1$ for ages i and older.

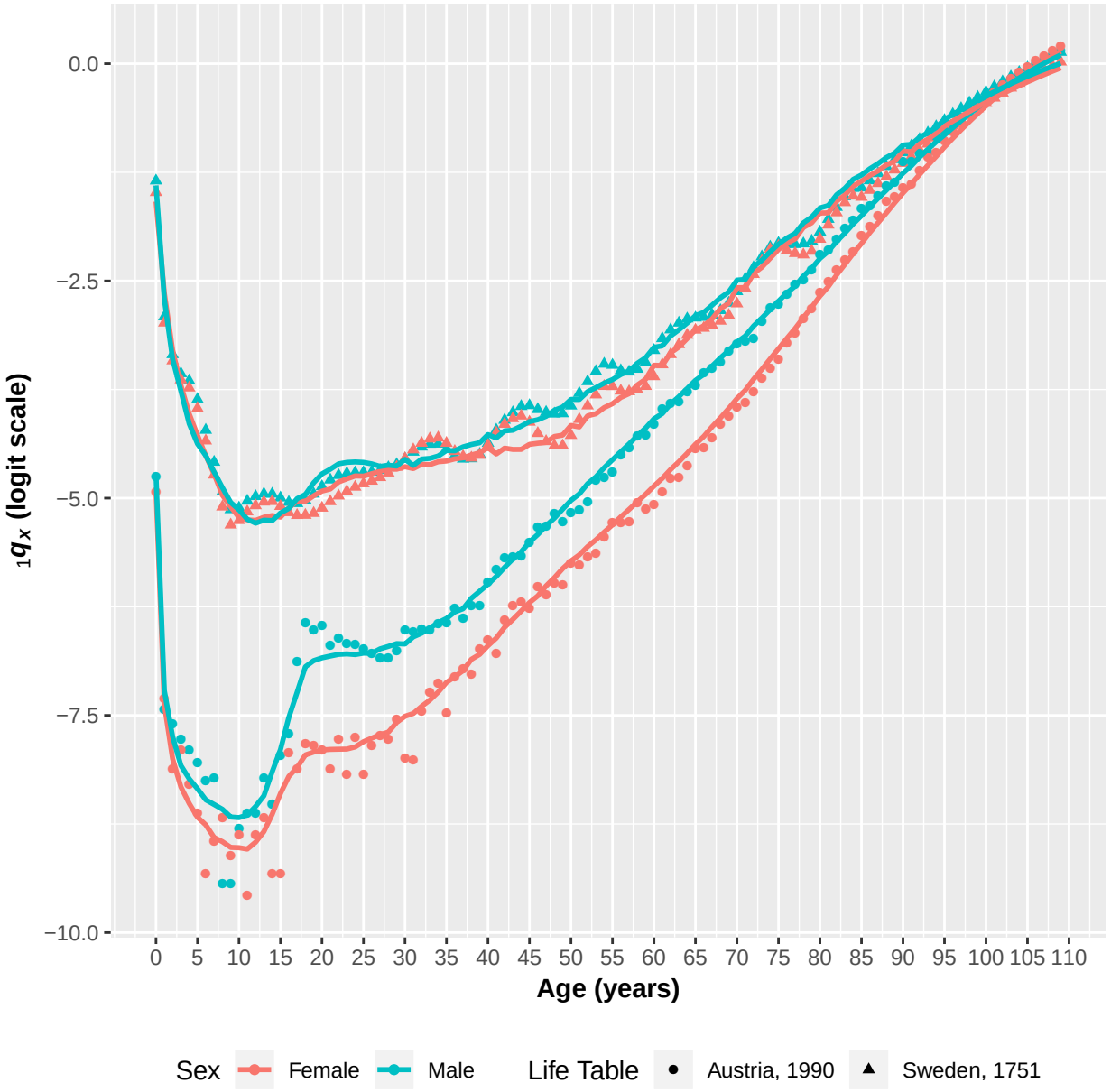

Figure 1: Example Data & Predictions. ${}_1q_x$ for very high mortality early in Sweden's time series and low mortality for a more recent year in Austria. Predicted values produced using ${}_5q_0$ alone as an input. Data as symbols and predicted values as lines.

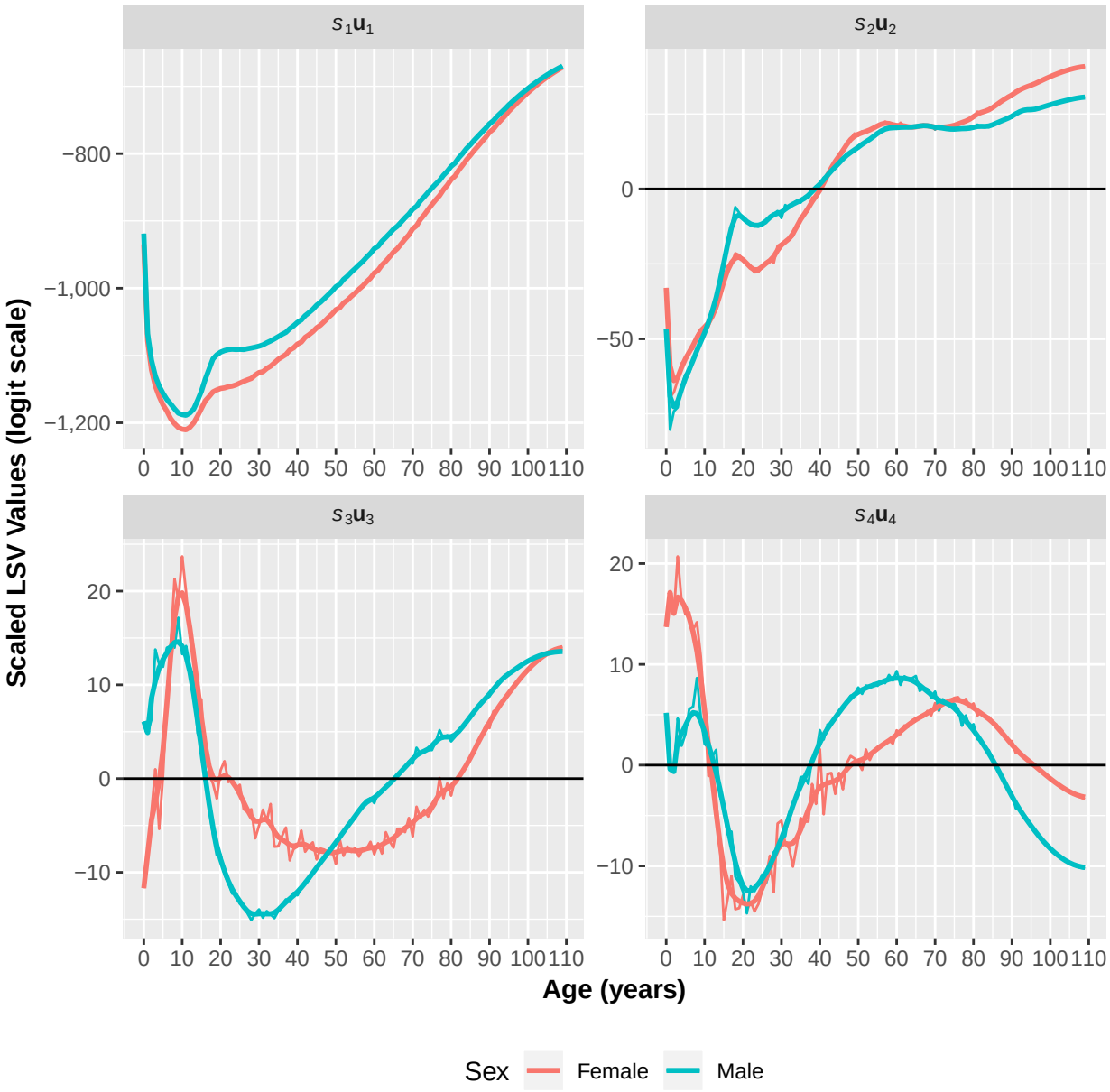

Figure 2: Scaled Left Singular Vectors. First four LSVs scaled by their corresponding SVs from the SVD of the 4,610 mortality schedules in the HMD. The more variable lines are *raw* components and less variable lines have been smoothed with a kernel smoother. The raw values are used throughout this work.

direction, especially at ages older than 90. However, the median error is never much more than 0.01, and as displayed in Figure 5, they are significantly smaller than the median errors produced by the Log-Quad model at the same ages. The error distributions of the in- and out-of-sample predictions are indistinguishable at all ages indicating that the SVD-Comp model is not sensitive to exactly which mortality schedules are used for calibration when half of them are used.

5.5 Varying Sample Size Cross Validation Prediction Errors

Figures 4 and E.6 in Appendix E (on-line appendices) contain the second set of cross validation results investigating the effect of varying the number of mortality schedules used to calibrate the SVD-Comp model. Both figures summarize the overall prediction error distributions (all ages and years combined) for the SVD-Comp model by sample status, in- versus out-of-sample mortality schedules. The sample fraction varies from 10% to 90% in increments of 20%. Figure 4 displays boxplots of the median of medians of overall error. This is very similar comparing in- and out-of-sample mortality schedules for both sexes across all sample fractions. There is a slight positive bias in all cases resulting from the positive bias in errors at older ages, see Figure 3. A similar situation exists for the distributions of the interquartile range of overall errors, Figure E.6 in Appendix E (on-line appendices). The only systematic change in these distributions by sample fraction is that the interquartile range of the indicators calculated from the sample decreases as the sample fraction increases, as expected. Inversely, there is a weak trend toward increases in the interquartile range calculated in the out-of-sample group as the sample fraction increases, also as expected. In general the SVD-Comp model appears to be remarkably robust as the number of mortality schedules used for calibration decreases. Performance is satisfactory all the way down to the 10% sample and good all all the way down to 30%.

5.6 Comparison between SVD-Comp and Log-Quad Prediction Errors

Figure 5 displays sex-age-specific boxplots of the distribution of prediction errors for both the SVD-Comp and Log-Quad models. The median error by sex and age is close to zero for both models through roughly age 70. At ages older than 70 the median error for the Log-Quad model is systematically substantially larger than zero, while for the SVD-Comp model the median error stays at zero. The sex-age-specific interquartile ranges are similar for both models, very small through roughly age 40, growing slowly between 40 and roughly 85 and then shrinking again through 110. In general at ages older than 45 the error distribution for the Log-Quad model is biased in a positive direction, while for the SVD-Comp model the error distribution is centered around zero at all ages.

Table 4 displays the total absolute errors on the natural scale for the SVD-Comp and Log-Quad models for predictions based on either ${}_5q_0$ alone or both $\left({}_5q_0, {}_{45}q_{15}\right)$. The table also presents differences between the total absolute errors for the two models in both additive (Log-Quad - SVD-Comp) and proportional form ([Log-Quad - SVD-Comp]/SVD-Comp). In all cases the SVD-Comp model predictions are globally closer to the the HMD life tables.

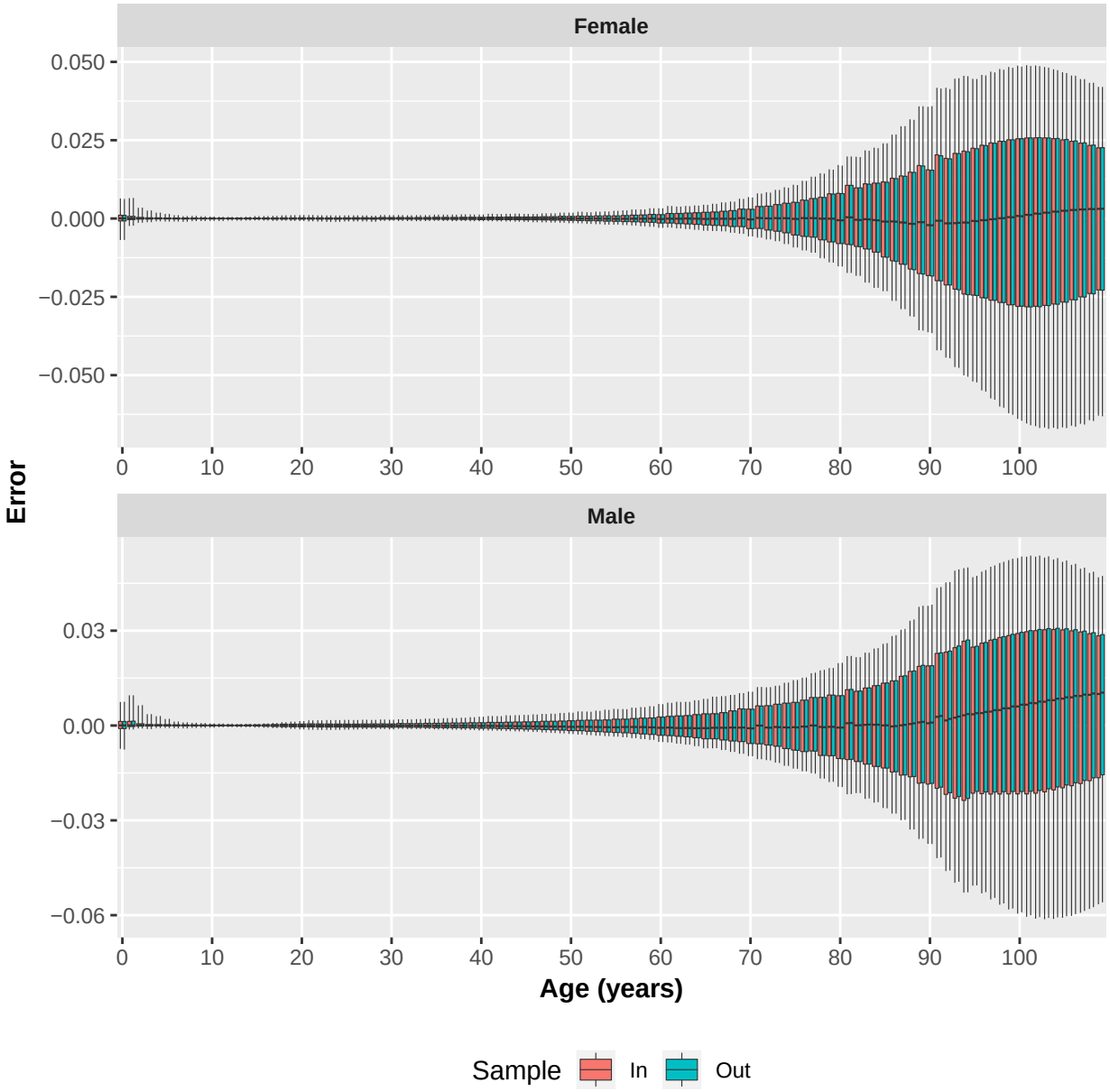

Figure 3: SVD-Comp Prediction Errors. Single-year age group prediction errors for in- and out-of-sample mortality schedules. 50 50% samples. Errors summarized over all in- and out-of-sample mortality schedules for the 50 samples. Whiskers extend to 10% and 90% quantiles.

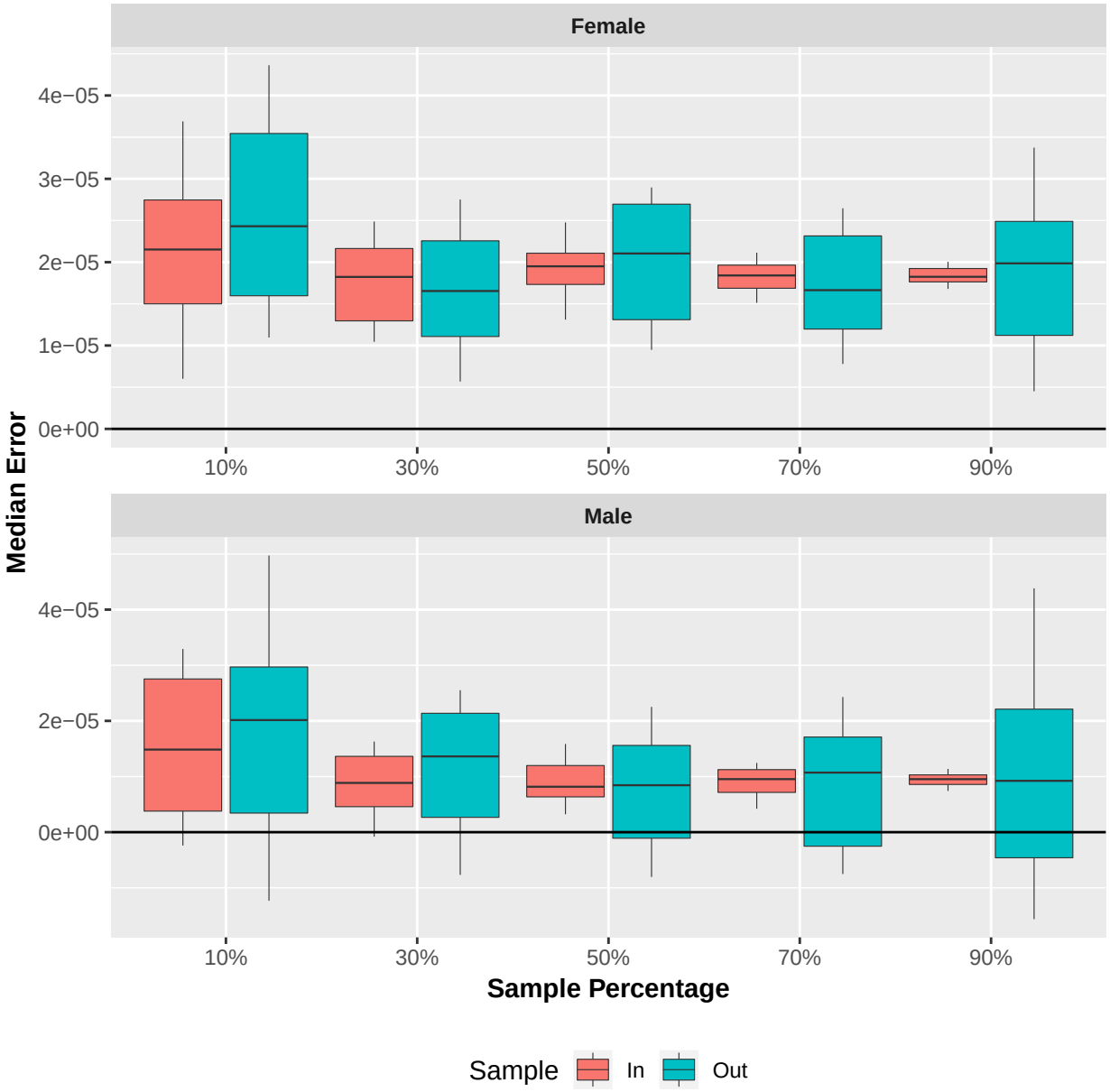

Figure 4: Median Prediction Error by Sample Fraction. 50 samples for each sample fraction. For each sample, median calculated across all ages and all mortality schedules in each sample category (in/out). Whiskers extend to 10% and 90% quantiles.

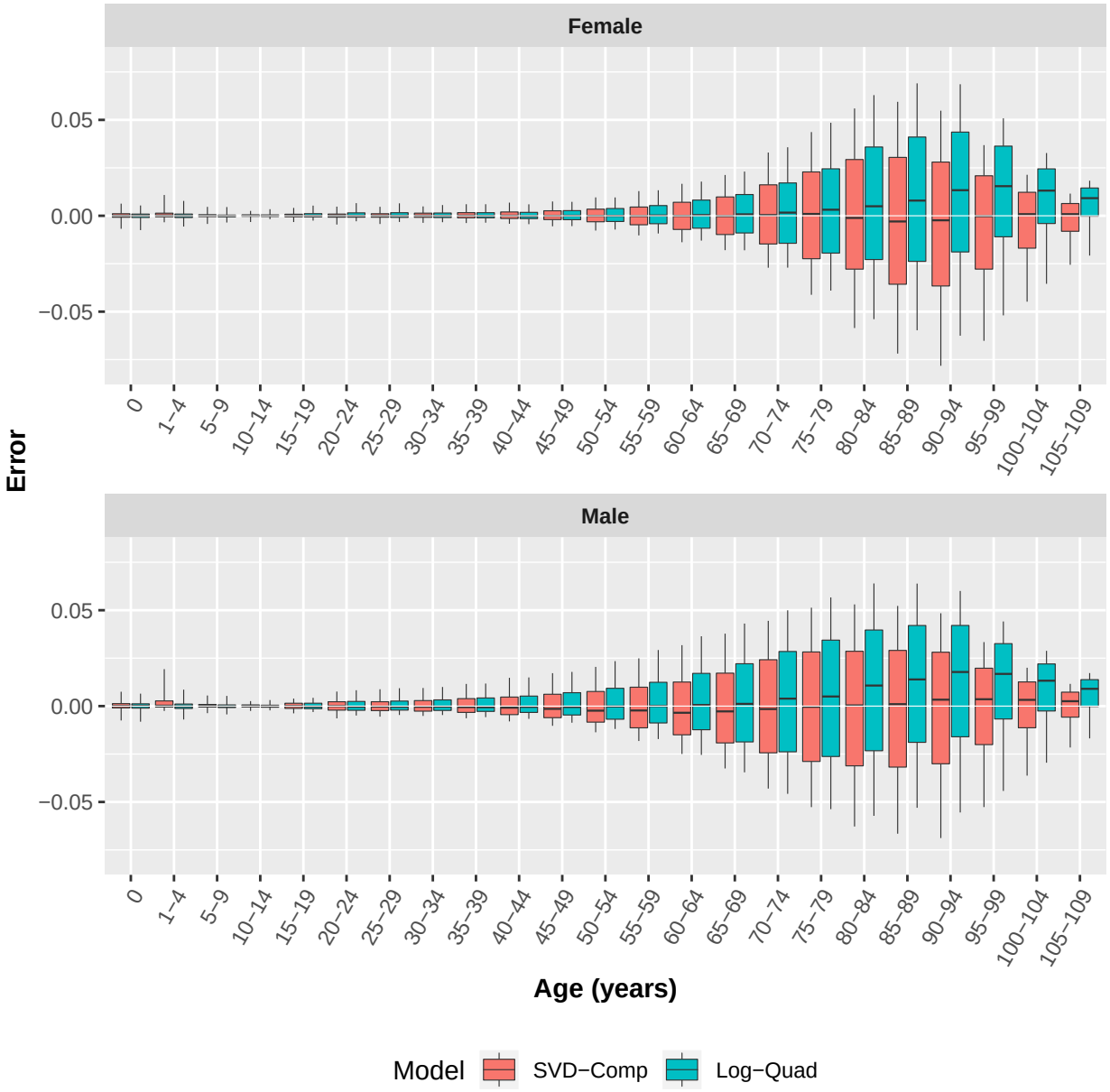

Figure 5: SVD-Comp and Log-Quad Prediction Errors. Five-year age group prediction errors for SVD-Comp and Log-Quad models using only child mortality ${}_5q_0$ as input. Whiskers extend to 10% and 90% quantiles.

Table 4: Summary of Prediction Errors for SVD-Comp and Log-Quad. Total absolute error and comparisons of total absolute error. Both models trained on all HMD life tables.

		Total Absolute Error Predicted by		
	Model / Summary	C1 ${}_5q_0$	C2 $({}_5q_0, {}_{45}q_{15})$	C3 C2-C1
Female				
R1	SVD-Comp	1,446	1,298	-148
R2	Log-Quad	1,502	1,399	-102
R3	R2-R1	56	102	46
R4	R3/R1 (%)	3.9	7.8	-30.9
Male				
R1	SVD-Comp	1,674	1,378	-296
R2	Log-Quad	1,777	1,472	-305
R3	R2-R1	103	94	-9
R4	R3/R1 (%)	6.1	6.8	3.0

Tables F.1 and F.2 in Appendix F (on-line appendices) display the weighted sum of age-specific absolute errors in $\widehat{\mathbf{q}}_\ell$ and $\widehat{\mathbf{e}}_\ell$ across all 4,610 life tables in the HMD. The last row in each displays the sum across all ages. The unweighted total absolute errors in $\widehat{e}_0$ for SVD-Comp calculated using one through four components are presented in Table F.3 in Appendix F (on-line appendices). Predicted values for life expectancy at birth, $\widehat{e}_0$, reflect predications at all ages so that errors in $\widehat{e}_0$ describe the cumulative effect of prediction errors at all ages. With each additional component, the total absolute errors in $\widehat{e}_0$ are reduced, and four components are required for SVD-Comp to perform better than Log-Quad. This is true in spite of the fact that the models used to predict the weights for the third and fourth components are not as predictive as those for the weights for the first two components (Equation 12 and Tables D.1 and D.2 in Appendix D (on-line appendices).

Finally, Figure E.5 in Appendix E (on-line appendices) displays predicted ${}_1q_x$ from the SVD-Comp using ${}_5q_0$ alone for three different levels of ${}_5q_0$.

5.7 Application to Mexico and South Africa

Figure 6 displays data and predictions from both Log-Quad and SVD-Comp in standard five-year age groups for Mexico 1983–1985 and South Africa 2005 using both child and adult mortality as predictors. The two models produce essentially the same predictions for Mexico, and both do an adequate job of following the data given that they are effectively two-parameter models. The situation for South Africa is different. As expected neither model is able to follow the HIV/AIDS-related bulge at adult ages. Both models do a reasonable job of threading the predictions through the male age schedule, overstating the mortality of adolescents and young adults and understating the mortality of middle-aged adults. For males the predictions are both plausible but unable to reproduce the bulge. SVD-Comp does the same for females, essentially cutting off the bulge, but

Log-Quad produces an implausible age-pattern of mortality with extremely high mortality for older children, adolescents, and young-to-middle-aged adults. The predictions for South Africa reveal a fundamental limitation of all empirically-based mortality models – the fact that they cannot represent mortality age profiles that are fundamentally different from those contained in the data used to create them. The solution to this is to identify or create new empirical life tables that represent the age profiles in question and include them in the data used to create the models.

6 Discussion

The SVD-Comp model is a simple framework for building mortality models. Its key advantages are 1) a simple linear structure that does not need to be changed to use the model in a variety of ways; 2) a general 'interface' through which input parameters can affect the age pattern of mortality, the weights in Equation 11; 3) an ability to handle new age groups without having to alter the fundamental structure of the model, including very short, like the one-year age groups used here; and finally 4) through its structure, an inherent constraint that ensures that mortality at each age is related to mortality at each other age according to the age patterns reflected in each of the components. Along with these, it also satisfies the combined list of desired characteristics for a mortality model enumerated in the introduction.

This approach is general and allows all-age (in arbitrarily fine age groups) mortality schedules to be predicted from any covariates that are related to age-specific mortality. This general relationship is quantified in the models (Equation 12) that relate the weights in Equation 11 to the covariates. Allowing this is the fact that the relationship of each age to all others is maintained through the constant components derived from the SVD, and those intra-age relationships are affected all together through the weights on the components. This constrains the intra-age relationships and relates them to the covariates in a simple, flexible way.

When the weights are modeled as functions of child mortality and calibrated using the relationship between the empirical weights ($v_{z\ell i}$ in Equation 8) and child mortality in the HMD, the model serves the same purpose as the Log-Quad (Wilmoth et al., 2012) model, and it performs slightly better in a direct comparison, while having the advantage of directly producing mortality schedules by single year of age. It is important to note that this comparison is conducted with the Log-Quad as presented in Wilmoth et al. (2012) and that in that article the authors explicitly favored an estimation technique that would, they claimed, reduce estimation bias at the cost of having (slightly) larger prediction errors when evaluated against the historical dataset, a fact that is apparent in Figure 5. The published Log-Quad was calibrated to the slightly different and smaller set of HMD life tables that existed at the time and met the authors' criteria for inclusion. Consequently the results of the comparison will likely change if the Log-Quad were recalibrated using the same set of HMD life tables described and used here. However, given how robust the SVD-Comp is to the set of life tables used in calibration (see Sections 5.4 and 5.5), this potential difference is unlikely to be large.

Concerning calibration and complexity, the cross validation results clearly demonstrate that the calibration to the HMD is robust with respect to exactly which and how many mortality schedules are used, and SVD-Comp is no more complex than Log-Quad. SVD-Comp requires one SVD calculation and six regression models (four in Equation 12, one in Equation 13, and one in Equation 14) for each sex to capture the relationship between child mortality and mortality at other ages in

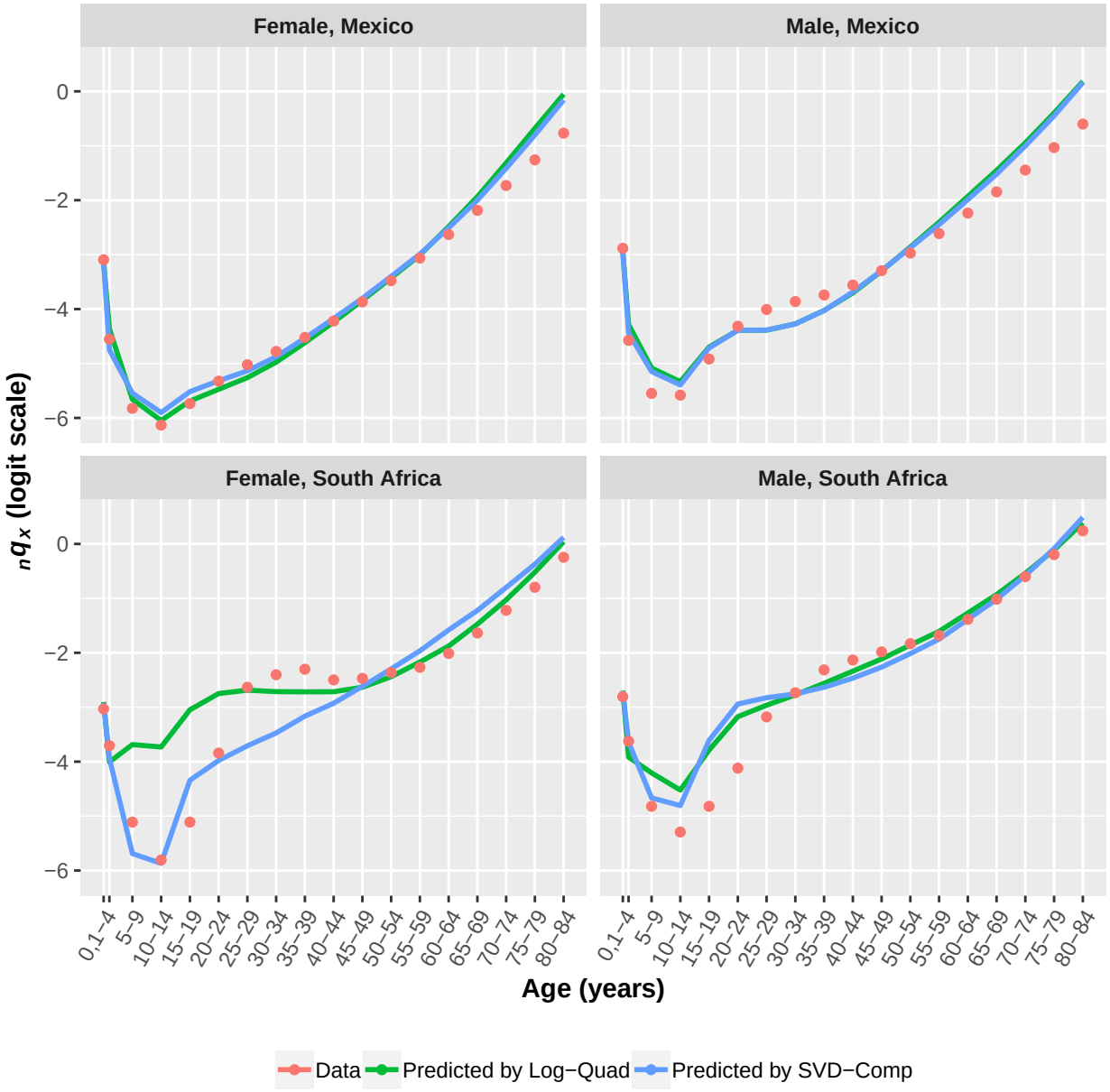

Figure 6: Application to Mexico and South Africa. Data and predicted values in standard five-year age groups produced by Log-Quad and SVD-Comp models using both child and adult mortality as predictors.

the HMD - 12 regression models in total. Log-Quad requires one SVD calculation and one log-quadratic model of the general form $\log({}_5m_x) \sim \log({}_5q_0) + \log({}_5q_0)^2$ for each five-year age group and another to refine the prediction of ${}_1q_0$ for each sex – 46 regression models in total. The total number of regression coefficients required by each model (for each sex) is: 44 for SVD-Comp and 70 for Log-Quad. The total number of discrete values required for prediction (for each sex) is: SVD-Comp - 484 (4.4/age group), and Log-Quad - 92 (3.8/age group). The models directly predict mortality in: SVD-Comp - single-year age groups, and Log Quad - five-year age groups. Comparing the complexity of the models is not easy and depends where one focuses, but it is clear that neither is obviously more/less complex than the other. Perhaps the only important difference in this respect is that there is nothing in the overall Log-Quad model to directly constrain the relationship of mortality at one age to another except for the quadratic form of the relationship between mortality at each age and ${}_5q_0$, whereas SVD-Comp manipulates a linear combination of age-specific vectors, so the relationships between ages are constrained to fall within the four-dimensional space defined by the four components used by SVD-Comp.

Together with our earlier work on an HIV-calibrated version of SVD-Comp (Sharrow et al., 2014), this demonstration suggests that it is reasonable to expect that SVD-Comp could be calibrated in a variety of additional ways to produce useful models that relate age-specific mortality to, for example, life expectancy at birth (or some other age), GDP, geographic region, time period, epidemiological indicators (as in Sharrow et al., 2014), a combination of any of these, or something else. Moreover, subtle effects on the age structure of mortality such as the 'rotation' in age-specific mortality identified by Li and Gerland (2011) could be incorporated by adding the necessary elements to the models for the weights. The same approach could be applied to develop models for the difference between underlying age-specific mortality and age-specific mortality affected by specific shocks such as natural disasters, conflict or epidemic disease such as HIV. It is even possible to refine the Lee-Carter model in Equation 1 by adding more components to the SVD-derived $\mathbf{b}_x k_t$ term so that the enhanced model could represent a wide range of age patterns instead of the constant age pattern included in the existing formulation. This would add more parameters to the model, but the payoff might be sufficient to make that worthwhile. Going further, the entire Lee-Carter model could be replaced by the SVD-Comp model which would give it the ability to model changing levels and age patterns of mortality independently and generally be more flexible.

The general SVD-Comp model in Equation 11 can be used in another way to interpolate or smooth incomplete or noisy age schedules by simply using OLS regression of the incomplete mortality schedule against the corresponding elements of the first few components $s_{zi}\mathbf{u}_{zi}$ with the constant constrained to be zero, and then predicting the full mortality schedule from all elements of the components and the coefficients estimated by the regression. Bayesian estimation can also be used to estimate the weights and their uncertainty, similar to Sharrow et al. (2013).

The application to Mexico and South Africa confirmed that the HMD-calibrated SVD-Comp works at least as well as Log-Quad when applied to mortality schedules in populations well outside of the HMD. For South Africa neither model was able to reproduce the HIV/AIDS-related mortality bulge at adult ages. SVD-Comp produced plausible mortality schedules for both sexes that were as close as possible to South Africa's, given that it could not reproduce the bulge. In contrast, Log-Quad produced a plausible mortality schedule for males but a nonsensical schedule for females. These results reveal an urgent need to increase the diversity of mortality schedules available in freely-accessible archives like HMD, and in particular, an important need to compile much better mortality data for Africa and other developing world regions where age schedules of mortality

are different from what has been observed in the developed world. Additionally, the South Africa application suggests that SVD-Comp may provide a stable framework to begin building mortality models that include epidemiological (e.g. HIV prevalence and ART coverage) and other predictors. Our earlier work using modeled data (Sharrow et al., 2014) is a start, but building models using modeled data is of limited value so we must assemble reasonable large, high quality empirical mortality data sets from the places where models such as Log-Quad and SVD-Comp are most useful.

7 Software & Reproducibility Materials

There is a GitHub repository containing all the code necessary to reproduce the results presented in this manuscript: `https://github.com/sinafala/svd-comp`. The on-line version of this article includes both the on-line appendices and a PDF rendered from the R Markdown file (on GitHub) that produces the results.

An R package (R Core Team, 2016) implementing the HMD child or child/adult mortality-calibrated version of SVD-Comp presented above is available as fully open source and free software to download directly from the GitHub repository using the *devtools* R package and command:

```
install_github(repo = "sinafala/svdComp5q0")
```

Acknowledgements

This work was supported in part by grant R01 HD054511 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). The funder had no part in the design, execution, or interpretation of the work. Tables of regression coefficients are formatted using the LaTeX package 'stargazer' (Hlavac, 2015).

References

Alexander, M., E. Zagheni and M. Barbieri. 2016. "A Flexible Bayesian Model for Estimating Subnational Mortality." *arXiv preprint arXiv:1607.03534* .

Bell, W. R. 1997. "Comparing and Assessing Time Series Methods for Forecasting Age-Specific Fertility and Mortality Rates." *Journal of Official Statistics* 13.

Bourgeois-Pichat, J. 1962. "Factor analysis and sex-age-specific death rates: a contribution to the study of the dimensions of mortality." *United Nations Population Bulletin* (6):147–201.

———. 1990. "Application de l'analyse factorielle à l'étude de la mortalité." *Population (french edition)* 45(4-5):773–802.

Bozik, J. E. and W. R. Bell. 1987. "Forecasting age specific fertility using principal components." In *Proceedings of the American Statistical Association, Social Statistics Section*, vol. 396, p. 401.

Brass, W. 1971. "On the scale of mortality." In *Biological Aspects of Demography*, edited by W. Brass, Taylor and Francis: London, UK, pp. 69–110.

Carter, L. R. and R. D. Lee. 1986. "Joint forecasts of US marital fertility, nuptiality, births, and marriages using time series models." *Journal of the American Statistical Association* 81(396):902–911.

Clark, S. J. 2001. *An Investigation into the Impact of HIV on Population Dynamics in Africa*. Ph.D. thesis, University of Pennsylvania.

———. 2015. "A Singular Value Decomposition-based Factorization and Parsimonious Component Model of Demographic Quantities Correlated by Age: Predicting Complete Demographic Age Schedules with Few Parameters." *arXiv preprint arXiv:1504.02057* .

Clark, S. J., M. Jasseh, S. Punpuing, E. Zulu, A. Bawah and O. Sankoh. 2009. "INDEPTH Model Life Tables 2.0." In *Annual Conference of the Population Association of America*, Population Association of America (PAA).

Clark, S. J. and D. J. Sharrow. 2011a. "Contemporary Model Life Tables for Developed Countries – An Application of Model-based Clustering." In *Annual Conference of the Population Association of America*, Population Association of America (PAA).

———. 2011b. "Contemporary Model Life Tables for Developed Countries: An Application of Model-based Clustering." *Center for Statistics and the Social Sciences (CSSS) Working Paper Series* (107). URL `http://www.csss.washington.edu/Papers/wp107.pdf`.

Coale, A. J. and P. Demeny. 1966. *Regional Model Life Tables and Stable Populations*. Princeton University Press.

Coale, A. J. and T. J. Trussell. 1974. "Model fertility schedules: variations in the age structure of childbearing in human populations." *Population Index* (1974):185–258.

Fosdick, B. K. and P. D. Hoff. 2012. "Separable factor analysis with applications to mortality data." *arXiv preprint arXiv:1211.3813* .

Golub, G. H., A. Hoffman and G. W. Stewart. 1987. "A generalization of the Eckart-Young-Mirsky matrix approximation theorem." *Linear Algebra and Its Applications* 88:317–327.

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

Good, I. J. 1969. “Some applications of the singular decomposition of a matrix.” *Technometrics* 11(4):823–831.

Heligman, L. and J. H. Pollard. 1980. “The age pattern of mortality.” *Journal of the Institute of Actuaries* 107(434):49–80.

Hlavac, M. 2015. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Harvard University, Cambridge, USA. URL `http://CRAN.R-project.org/package=stargazer`, r package version 5.2.

INDEPTH Network. 2002. *INDEPTH Mortality Patterns for Africa*, *Population and Health in Developing Countries*, vol. 1, chapter 7. Ottawa: IDRC Press, pp. 83–128.

Ledermann, S. 1969. “Nouvelles Tables-types de Mortalité.” No. 53 in INED Traveaux et Documents, Paris: Presses Universitaires de France.

Ledermann, S. and J. Breas. 1959. “Les dimensions de la mortalité.” *Population* 14(4):637–682.

Lee, R. D. 1993. “Modeling and forecasting the time series of US fertility: Age distribution, range, and ultimate level.” *International Journal of Forecasting* 9(2):187–202.

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

Li, N. 2015. “Estimating Life Tables for Developing Countries.” Tech. Rep. 2014/4, United Nations Department of Economic and Social Affairs Population Division, `http://www.un.org/en/development/desa/population/publications/pdf/technical/TP2014-4.pdf`.

Li, N. and P. Gerland. 2011. “Modifying the Lee-Carter Method to Project Mortality Changes up to 2100.” Paper presented at the 2011 Annual Meeting of the Population Association of America (PAA), Washington, D.C., March 31-April 2.

Li, T. and J. J. Anderson. 2009. “The vitality model: A way to understand population survival and demographic heterogeneity.” *Theoretical Population Biology* 76(2):118–131.

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

Max Planck Institute for Demographic Research, University of California, Berkeley and The Institut d’études démographiques (INED). Downloaded August 2018. *Human Life Table Database*. `https://www.lifetable.de/data/hld.zip`.

Murray, C. J., B. D. Ferguson, A. D. Lopez, M. Guillot, J. A. Salomon and O. Ahmad. 2003. “Modified logit life table system: principles, empirical validation, and application.” *Population Studies* 57(2):165–182.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

R Foundation for Statistical Computing. 2016. *The R Project for Statistical Computing*. http://www.r-project.org. URL http://www.R-project.org.

Sharrow, D., S. J. Clark, M. Collinson, K. Kahn and S. Tollman. 2013. "The age pattern of increases in mortality affected by HIV: Bayesian fit of the Heligman-Pollard Model to data from the Agincourt HDSS field site in rural northeast South Africa." *Demographic Research* 29(39):1039–1096. URL http://www.demographic-research.org/volumes/vol29/39/.

Sharrow, D. J., S. J. Clark and A. E. Raftery. 2014. "Modeling age-specific mortality for countries with generalized HIV epidemics." *PloS ONE* 9(5):e96447.

Stewart, G. W. 1993. "On the early history of the singular value decomposition." *SIAM review* 35(4):551–566.

Strang, G. 2009. *Introduction to Linear Algebra 4e*. Wellesley-Cambridge Press.

United Nations, Department of Economic and Social Affairs, Population Division. 1955. *Age and Sex Patterns of Mortality: Model Life-tables for Under-developed Countries*. New York: United Nations Department of International Economic and Social Affairs Population Division.

———. 1982. *Model life tables for developing countries*. No. 77, New York: United Nations Department of International Economic and Social Affairs Population Division.

———. 2015a. "File 0-2: Latest data sources used to derive estimates for total population, fertility, mortality and migration by countries or areas in WPP 2015 revision: POP/DB/WPP/Rev.2015/F0-2." https://esa.un.org/unpd/wpp/DVD/Files/4_Other\%20Files/WPP2015_F02_METAINFO.XLS.

———. 2015b. *World Population Prospects: the 2015 Revision*. New York: United Nations.

———. 2015c. *World Population Prospects: The 2015 Revision, Methodology of the United Nations Population Estimates and Projections*. Working paper No. ESA/P/WP.242.

University of California, Berkeley and Max Planck Institute for Demographic Research. *Human Mortality Database*. http://www.mortality.org or http://www.humanmortality.de.

Wang, H., L. Dwyer-Lindgren, K. T. Lofgren, J. K. Rajaratnam, J. R. Marcus, A. Levin-Rector, C. E. Levitz, A. D. Lopez and C. J. L. Murray. 2013. "Age-specific and sex-specific mortality in 187 countries, 1970–2010: a systematic analysis for the global burden of disease study 2010." *The Lancet* 380(9859):2071–2094.

Wilmoth, J., J. Vallin and G. Caselli. 1989. "Quand certaines générations ont une mortalité différente de celle que l'on pourrait attendre." *Population* 44(2):335–376.

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

Wilmoth, J. R. 1988. *On the Statistical Analysis of Large Arrays of Demographic Rates*. Ph.D. thesis, Department of Statistics, Princeton University.

———. 1990. "Variation in Vital Rates by Age, Period, and Cohort." *Sociological Methodology* 20:295–335.

Wilmoth, J. R. and G. Caselli. 1987. “A simple model for the statistical analysis of large arrays of mortality data: rectangular vs. diagonal structure.” *IIASA Working Paper* (WP-87-058).

World Health Organization. Downloaded August 21, 2018. *Global Health Observatory data repository*. `http://apps.who.int/gho/data/?theme=main&vid=61540`.

Zaba, B. 1979. “The four-parameter logit life table system.” *Population Studies* 33(1):79–100.