Modeling the Age Pattern of Human Mortality: Mathematical and Tabular Representations of the Risk of Death

David J. Sharrow

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Samuel J. Clark, Chair

Adrian E. Raftery

Jerald R. Herting

Program Authorized to Offer Degree:
Sociology

University of Washington

Abstract

Modeling the Age Pattern of Human Mortality:
Mathematical and Tabular Representations
of the Risk of Death

David J. Sharrow

Chair of the Supervisory Committee:
Associate Professor Samuel J. Clark
Sociology

The age-pattern of vital events is one of the oldest and most well studied topics in demography. Mortality is no exception with the introduction of the life table as a central demographic tool in the mid 17th century. Over the past 50 years, population analysts have developed mathematical equations to parsimoniously describe the shape of human mortality across age allowing for succinct time-series comparisons, but many of these models can only describe restricted portions of the age range and at times can be difficult to fit due to the necessarily large number of parameters needed to describe the age pattern of mortality. Furthermore, owing to estimation difficulties or data limitations, these models are rarely applied to African populations. The first part of this dissertation is devoted to investigating the sex-age-specific changes in mortality in a prospectively monitored rural population in South Africa during the course of the HIV epidemic. These changes are quantified by fitting the eight parameter Heligman-Pollard model of age-specific mortality using a robust Bayesian estimation method. Fitting a model like the Heligman-Pollard or calculating summary mortality indicators such as life expectancy at birth, requires a complete set of age-specific mortality rates. Absent high-quality vital registration systems that could collect age-specific mortality data, many countries rely on indirect techniques like *model life tables* – models that produce a complete set of mortality rates from age restricted mortality indicators such as the probability of death before age five. Existing model life table systems are outdated and unable to produce certain mortality patterns like those with very low levels of child mortality or high adult mortality resulting from

HIV. The second portion of the dissertation is devoted to generating a model life table system for both high and low-to-middle income countries by identifying similar age patterns of mortality in a collection of 844 life tables from the Human Mortality Database and 329 life tables from the INDEPTH network of Health and Demographic Surveillance Sites spread throughout Africa and parts of Asia.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work benefited greatly from the contributions of several people and organizations. I am particularly grateful to the International Network for the Demographic Evaluation of Populations and Their Health (INDEPTH) as well as the Agincourt HDSS for the generous use of data and other resources during the completion of this work. I wish to express my appreciation to all members of my dissertation committee with special thanks to Adrian Raftery whose generous support and insightful remarks contributed greatly to my training and to Sam Clark for his years of warm encouragement, earnest advice, constructive comments, and all that funding.

DEDICATION

To Anita Van Buren, Claire Kincaid, Jamie Ross, Abbie Carmichael, Nora Lewin, Serena Southerlyn, Alexandra Borja, Connie Rubirosa, and Nina Cassady

Chapter 1
INTRODUCTION

The age pattern of vital events is a historically important topic in demography and related fields. The age pattern of mortality is no exception with attempts to capture the age variation in the risk of death dating back to the establishment of the life table, which describes the dying out of a cohort across age, as an important demographic tool in the mid-seventeenth century (Graunt, 1662; Halley, 1693; Deparcieux, 1746). Since that time, several alternative and often more parsimonious means of characterizing the age pattern of mortality have emerged. Monitoring the age pattern of mortality or other vital events is a nontrivial pursuit when one considers the tremendous impacts these patterns can have on the social, economic, and political characteristics of a society especially when limited resources are available to meet these pressures. Continuing this traditional line of research, the focus of this dissertation is modeling and quantifying the age pattern of human mortality. I explore this topic by building on an existing mathematical model of mortality and developing model life table systems for both high and low-to-middle income countries.

This chapter introduces the motivation, data, and, briefly, the methods described in more detail in the upcoming chapters as well as the overall organization of the dissertation. For the most part each chapter is meant to be a self-contained work which one could read on its own without any additional material although chapters three and four use a similar method, which is detailed in chapter three with a summary in chapter four. The final chapter summarizes and contextualizes the findings and implications of this work, discusses some of the limitations, and describes future lines of research.

1.1 Modeling the Age Pattern of Mortality

Although the epidemiological profile of a population largely governs the shape of age-specific mortality, the human mortality pattern contains several common features found in many populations. Typically the mortality rate declines sharply over the first few years

of life and begins an exponential increase past the middle ages. Women often see a slight bulge in the young adult years resulting from maternal mortality while males experience a slightly higher young adult mortality hump due to accidents during this period of the life span. Recently, as life spans take on a more protracted form, we see mortality rates increasing but at a decreasing rate at the oldest ages (80+). The life table works well to disaggregate overall mortality into narrower age intervals revealing the human mortality pattern, but the large set of numbers (anywhere from 19 to 24 depending on the number of age intervals) can be unwieldy especially for comparisons over time. Thus, analysts have sought to establish mathematical functions or identify commonly observed age-patterns that can represent the shape of age-specific mortality with as few parameters as possible.

The advantages of parsimonious representation of the risk of death over age are numerous for demographers. Parametric functions often have the convenient property of demographically interpretable parameters, which easily summarize and decompose mortality over certain portions of the age-range making comparisons over time possible (Roger, 1987; Hartmann, 1987; McNown and Rogers, 1989; Debón et al., 2005). The analyst may also use a model of age-specific mortality to smooth noisy data or obtain mortality rates for age groups not contained in the available data. Demographers use models that identify regularities in the age pattern of mortality among a large collection of life tables to estimate historical mortality patterns when data is unavailable–an essential component to certain population projection methods (Roger, 1987; McNown and Rogers, 1989; Preston et al., 2001).

One could divide the approaches to modeling the age pattern of mortality into two broad categories: mathematical and tabular.[1] Mathematical models, often referred to as 'laws of mortality,' posit a mathematical function where mortality rates are a function of age. Tabular representations, which include model life tables, present full life tables based on a model pattern and indexed according to the level of mortality (often by life expectancy at birth).

[1]Preston et al. (2001) identify a third category, relational models, which combines mathematical and tabular models where a mathematical function is formulated to relate predicted rates to some standard set of rates. A relational model is not developed in this dissertation so I restrict the discussion to the two broader categories.

1.1.1 Mathematical Representations: 'Laws of Mortality'

Capturing the shape of the mortality curve in a mathematical equation is necessarily complex and can involve many parameters although fewer than the 19 or so age groups typically used in period life tables. The history of mathematical models is marked by added complexity as analysts attempted to craft a single law of mortality that could represent the full range of human mortality experiences. Early on Gompertz (1825) proposed a two parameter model that reproduces the exponential increase in mortality risk with age past the middle ages, but Gompertz's law was only meant to represent 'underlying' mortality–mortality that does not include accidents or infectious disease (Preston et al., 2001). Makeham (1860) later built on Gompertz's model by adding a parameter to capture these additional sources of mortality, but neither of these models represents childhood mortality well or old-age mortality where rates continue rising but at a decelerating rate (Vaupel et al., 1979; Horiuchi and Coale, 1982; Preston et al., 2001). Since the introduction of these early models, demographers and actuaries have advanced several other 'laws' to capture additional portions of the age range including a logistic model by Perks (1932) to capture the slower rate of change in mortality above age 80 and the five parameter Siler (1979) model of all-age mortality, which was shown to fit well to American and Australian historical patterns (Gage and Mode, 1993). Heligman and Pollard (1980) proposed an eight-parameter model that decomposes mortality into three additive terms – a child mortality portion to capture the steep decline in mortality at the earliest ages, a flexible adult mortality 'hump' originally designed to fit accident mortality in males and, to a lesser extent, maternal mortality in females, and an old-age mortality term similar to the Gompertz and Perks models. The additional parameters result in a flexible model that can represent a wide range of mortality patterns although its use has been restricted to data from developed countries.

Although a more parsimonious description of the shape of human mortality, these 'laws' are not without drawbacks. Because the shape of mortality risk across the age range is determined by an epidemiological profile, capturing the variety of possible shapes in a single law for all populations is impracticable. For example, although the early models mentioned above do well to capture many mortality experiences, the HIV/AIDS epidemic in certain countries has caused unprecedented reversals in life expectancy as a

result of a dramatic increases in deaths among working-ages adults (Dorrington et al., 2001). These same populations are dealing with killing mechanisms at other portions of the age range as well such as noncommunicable diseases in old age, a situation referred to as a 'double burden' of disease (Marshall, 2004). The disease burden in these regions constitutes a partial reversal of the epidemiological transition as described by Omran (1971) where countries move from high mortality resulting from infectious and parasitic disease to low mortality with deaths concentrated in old age from chronic illness. The combination of these unusual mortality conditions requires a highly parameterized model with a flexible adult mortality hump like the Heligman-Pollard to capture this shape. The increase in complexity with the introduction of additional parameters can lead to an over-parameterized (non-identified) model where numerous parameter value combinations will result in similar goodness of fit.

Chapter two of this dissertation fits the Heligman-Pollard mortality model to data from rural South Africa over a 14 year period with a Bayesian estimation procedure. This work has two aims.

1. Quantify changes in the sex-age pattern of mortality experienced by a population with endemic HIV

2. Demonstrate the use of a robust Bayesian estimation method for the HP model that accounts for uncertainty

The parameters of the HP model are fully demographically interpretable and I relate the changes in parameter estimates to the change in lagged HIV prevalence to describe the likely impact of HIV on age-specific, all-cause mortality during the course of the epidemic.

The data for chapter two are the age-specific death counts and person-years collected between 1994 and 2007 from the Agincourt Health and Demographic Surveillance Site (HDSS) located in rural northeast South Africa (Kahn et al., 2007; INDEPTH, 2011). The age-specific probabilities of death are presented in figure 1.1. Located in Mpumaglunga province, which has seen HIV prevalence jump from 1.3% at the beginning of the 1990s to near 30% just a decade later, Agincourt has experienced major heath transitions including increasing child and adult mortality (Kahn et al., 2007). These changes are apparent in

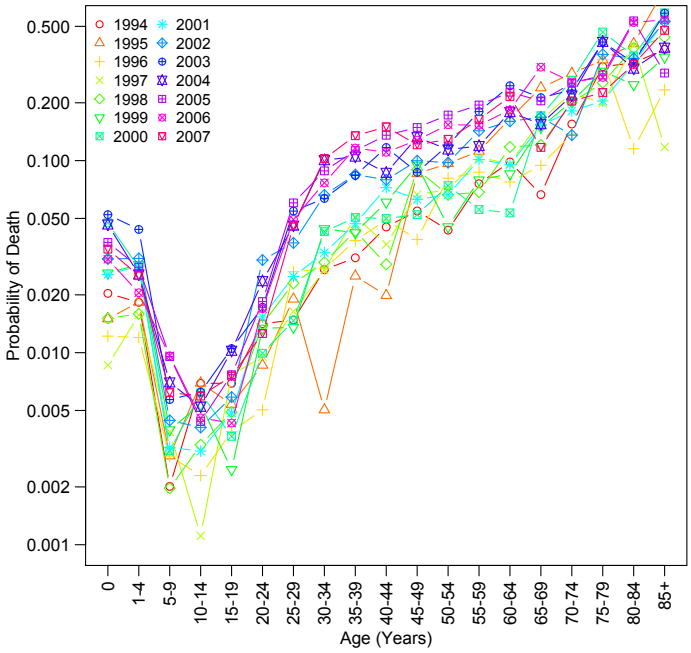
figure 1.1 especially around the turn of the century when HIV related mortality begins to influence the mortality curve substantially.

The Bayesian fitting method has dual advantages. The first addresses the over-parameterization and fit issues described above, which are well documented for the Heligman-Pollard (Congdon, 1993; Dellaportas, 2001; Bebbington et al., 2007). Moreover, the Bayesian approach yields probability distributions of the model parameters that can be used to produce probability distributions for any life table quantity. One can make inference from these distributions and determine if and when this population experienced significant declines in life expectancy during the period under study.

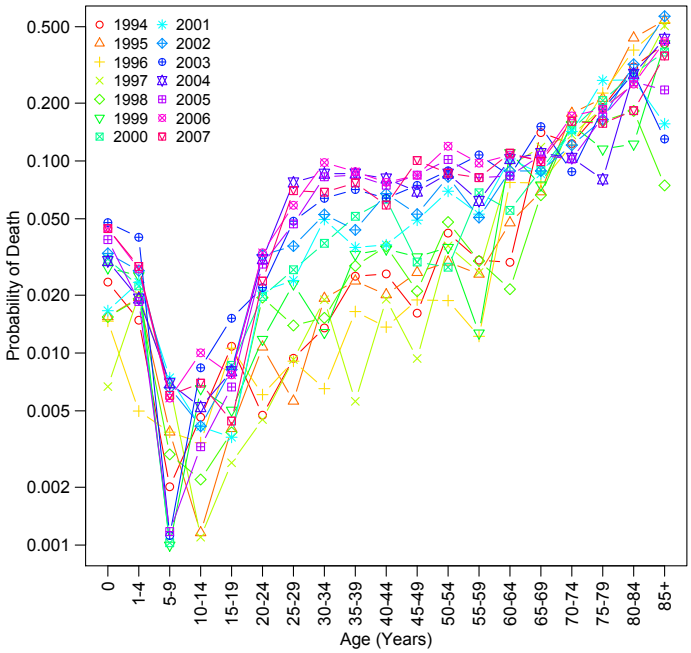
1.1.2 Tabular Representation: Model Life Tables

In order to fit a parametric function like the Heligman-Pollard model, one needs a complete set of age-specific mortality rates. Likewise, certain summary mortality indicators, such as life expectancy at birth, reported by numerous governments and non-governmental organizations concerned with population dynamics and wellbeing, are derived from life tables calculated from a complete set of age-specific mortality rates. Gold standard for collecting this kind of data would be complete civil vital registration of birth and death events, but countries without adequate vital registration systems must rely on indirect methods like model life table systems to estimate age-specific mortality. Model life table systems represent typical age patterns of mortality that can be used for many purposes aside from their primary purpose of extrapolating full age patterns of mortality from age-restricted (but often more widely available) indicators such as the probability of death between birth and age 5. These models can be used to smooth noisy data, obtain mortality rate estimates for age intervals not contained in the available data, or to obtain historical mortality patterns for the purpose of population projection.

Data availability and accuracy, and thus the necessity for model life tables, varies greatly across countries. Roughly half the countries in the world do not have at least a partial vital registration system that tracks age-specific mortality while approximately 39% of all countries (≈ 90) have some information on at least child mortality from which a compete set of age-specific mortality rates can be estimated using model life tables (Gerland, 2009). Limited data availability and reliability are, of course, not randomly



(a) Male



(b) Female

Figure 1.1: The Age Pattern of the Probability of Death for Agincourt HDSS 1994-2007

distributed across region. Nearly all countries in Europe and North America have at least a partial vital registration system while data from low- and middle-income countries remains limited (Mathers et al., 2005; Gerland, 2009).

Model life table systems are built on large collections of observed life tables where typical patterns and regularities can be identified. Once a model mortality pattern is established, the level of mortality is then indexed by life expectancy. Because these models are built of collections of empirical life tables, they can only represent mortality patterns for the regions and time periods contained in the databases on which they were built. Current model life table systems were generated using data from restricted geographical locations and time periods, with the most widely used systems dating back some 30 years. These systems are unable to reflect contemporary mortality experiences including high adult mortality resulting from causes such as HIV or the extremely low childhood mortality observed in some contemporary developed world settings (Coale and Guo, 1989; Wilmoth et al., 2009). A review of existing and widely-used model life table systems follows.

Widely Used Model Life Table Systems Perhaps the most frequently used model life table system is that of Coale and Demeny (1966, second edition 1983). Beginning with a collection of 326 life tables for each sex, Coale and Demeny dropped life tables that were not derived from registration data or complete enumeration of the population to which they refer. Each table was then compared to a composite table at a similar level of mortality and visual examination was used to screen tables with large deviations. The system was built on the remaining 192 life tables by sex. The final set of tables contained only 16 tables from Asia and Africa. The remaining 176 came from North America, Europe, Australia, and New Zealand. Four regional patterns were identified in the 192 tables: North, South, East, and West, which is considered the most general. For each of these regions, the level of mortality is indexed by life expectancy at birth. In many instances, the Coale and Demeny models work well, but as Preston et al. (2001) point out, the result is sensitive to the choice of a regional model. This model, like all model life table systems, cannot fit mortality conditions that lie outside the scope of the original data, which includes low child mortality and high adult mortality.

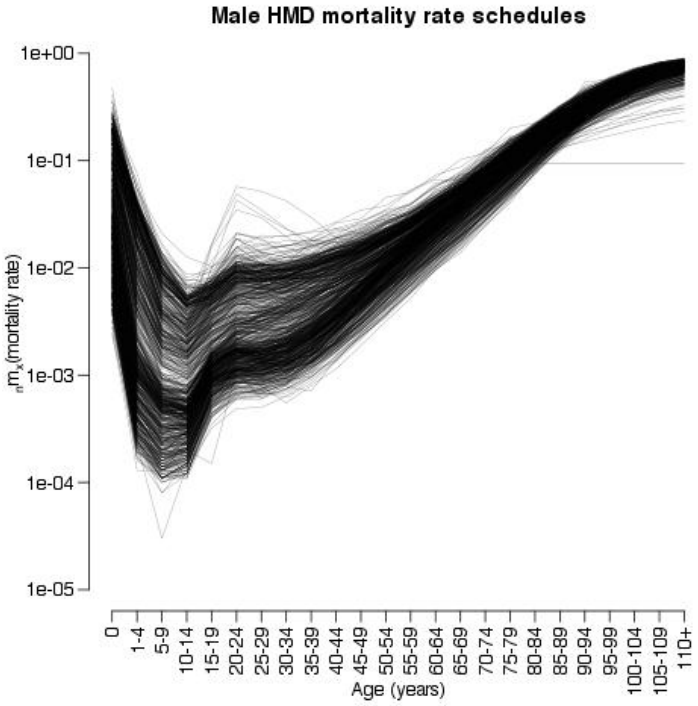
Using a combination of several previous approaches like graphical examination, selec-

tion into regional clusters, and principal components analysis the United Nations (1982) published a set of model life tables for developing countries. After rigorous data quality checks only 72 of the original 286 tables remained (with only data from Tunisia to represent Africa) from which five clusters were identified. Again, four regional clusters emerged (Latin American, Chilean, South Asian, Far Eastern) as well as a fifth 'General' pattern based on all 72 tables, for which mortality levels were indexed by life expectancy. The UN system did address the lack of information from the developing world where model life tables are perhaps most relevant, but this model ultimately lacks sufficient data from Africa.

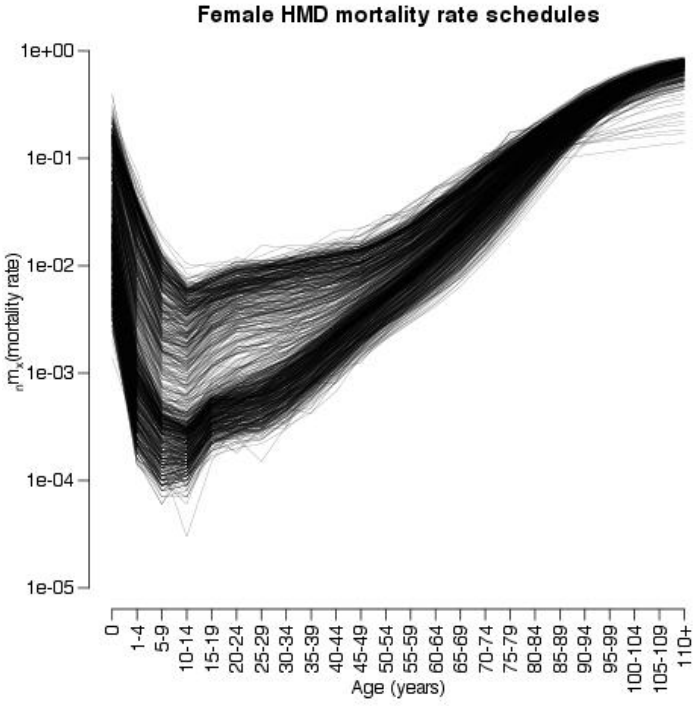
Although not technically a tabular model, the WHO modified logit model is, in effect, a model life table system (Murray et al., 2003). Also based on a large collection of empirical, historical life tables ($\approx$ 1800 including about a third originating from developing countries), this approach modifies the two-parameter logit system proposed by Brass (1971) that posits a mathematical relationship to relate a pair of tables one of which is a standard.[2] Because this model is based on such a large collection of life tables, it performs well for a number of countries and mortality experiences but still lacks satisfactory data from Africa with just one table from Tunisia (1968), three from South Africa (1941, 1951, 1960), and nine from Mauritius (1990-98). Another newly advanced and remarkable accurate model is that of Wilmoth and colleagues (Wilmoth et al., 2012) based on 719 life tables from the Human Mortality Database, a large and well maintained collection of life tables spanning three centuries and dozens of mostly developed countries. Wilmoth's model, like the WHO system, is two dimensional but with the second input being optional. Also like the WHO model, both input parameters are continuous removing the family/level structure of the Coale and Demeny and UN model life tables. As with all other model life table systems, Wilmoth's does not include any data from Africa and very little from the developing world (although one could re-estimate his model with these kind of data should they become available).

Building on earlier work (Clark, 2002; Clark et al., 2009), chapters three and four describe an easy-to-use model life table system based on observed mortality patterns for high income and low-to-middle income countries respectively. For a model life table system

[2]Although any standard can be chosen, choice of this standard along with the equation which relates pairs of tables determines the accuracy of this method (Preston et al., 2001).

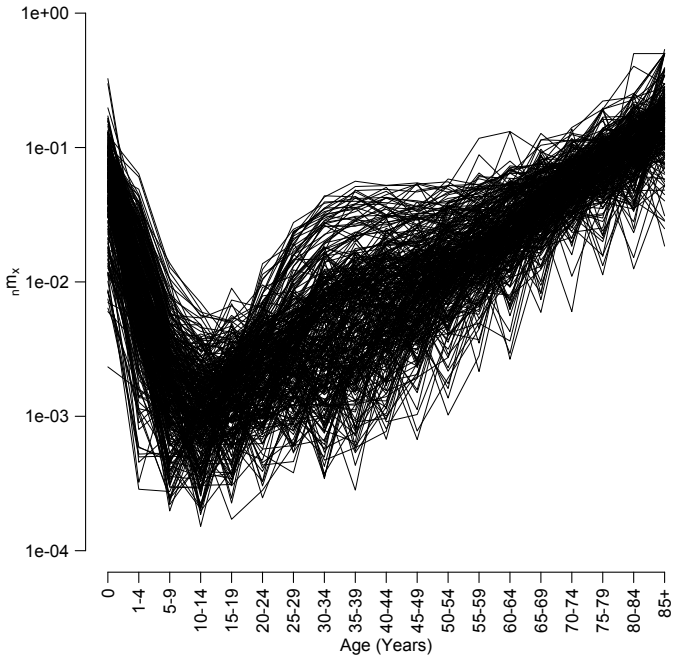


(a) Male

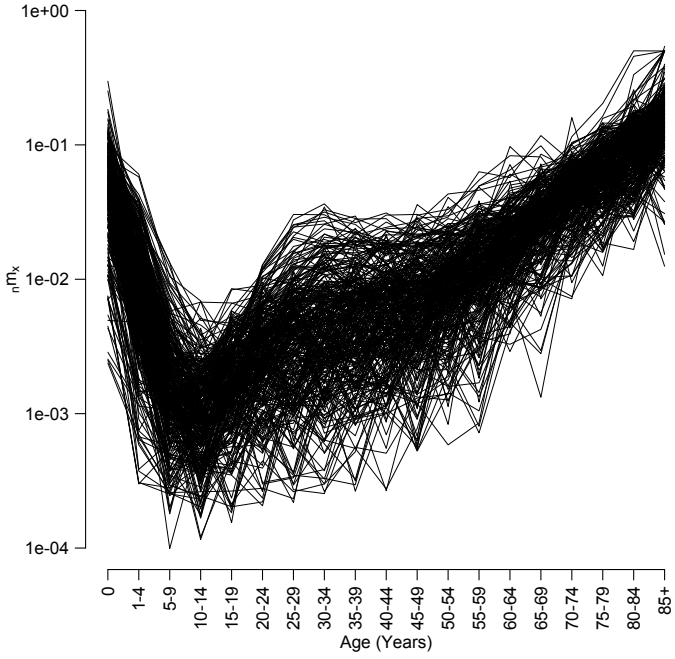


(b) Female

Figure 1.2: The Age Pattern of Mortality Rates for Life Tables in the Human Mortality Database (n=844). Each line in this figure represented one life table from the HMD collection.



(a) Male



(b) Female

Figure 1.3: The Age Pattern of Mortality Rates for Life Tables in the INDEPTH database (n=329). Each line in this figure is one life table from the INDEPTH collection.

for high-income countries, I identify similar age patterns of mortality in a collection of 844 life tables from the Human Mortality Database (HMD) (University of California, Berkeley and Max Planck Institute for Demographic Research, Data Downloaded November 2009) using a model based clustering method (Fraley and Raftery, 2006). The HMD data represent 37 mostly developed countries. The earliest tables date back to the mid 18th century and the most recent from 2007. The HMD is a high quality, well documented, publicly available source of mortality data making it ideal to develop this method of life table construction.

In chapter four, I develop a model life table system for low-to-middle income countries using a collection of life tables from the International Network for the Demographic Evaluation of Populations and Their Health (INDEPTH), a network of health and demographic surveillance sites (HDSS) found throughout Africa and southeast Asia (INDEPTH, 2011). Each HDSS longitudinally monitors defined subjects like individuals or households within a defined geographic area. Data for chapter four come from 32 HDSS of variable sized populations and periods of coverage from 1983 to 2011. This collection offers a unique opportunity to include data from Africa as well as life tables form regions affected by severe HIV epidemics, which is not well represented in any life table system to date. The collection of mortality rate schedules from the HMD and INDEPTH are shown in figures 1.2 and 1.3. Given that there is sufficient variation in the age pattern of mortality contained within these datasets, a model life table system based on them ought to be able to represent a wide array of human mortality experiences.

There are two aims for each of the chapters just described.

1. Develop, demonstrate, and validate the use of a model life table system for high and low-to-middle income countries based respectively on the Human Mortality Database and INDEPTH collections of life tables

2. Describe the temporal and regional patterns of mortality contained within these two datasets

Chapter 2

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

2.1 Introduction

The work in this chapter makes two contributions. *The first is a detailed description of the likely impact of HIV on period-sex-age-specific, all-cause mortality as the HIV epidemic grows in a population living in rural northeast South Africa.* This task is accomplished by reinterpreting the parameters of the eight-parameter Heligman-Pollard mortality model (Heligman and Pollard, 1980; Rogers and Gard, 1991; Gage and Mode, 1993; Congdon, 1993) and fitting it to age-specific probabilities of dying. *The second is a Bayesian fitting procedure developed to produce robust fits of the Heligman-Pollard mortality model that yield probability distributions for the parameters, the age-specific probabilities of dying, and the other columns of the corresponding life tables.*

Populations with high HIV prevalence in east and southern Africa have experienced dramatic increases in mortality (Dorrington et al., 2001; UNAIDS and WHO, 2009) and corresponding declines in life expectancy (UNAIDS and WHO, 2008). Because the primary modes of transmission of HIV in those populations are heterosexual sex and mother-to-child transmission at birth or through breastfeeding, the majority of HIV positive people are either very young children or sexually-active adults. Consequently there is a characteristic age pattern of deaths resulting from AIDS: very young children who progress through the disease quickly and middle-aged adults who are HIV positive for roughly ten years before dying of AIDS. Adult women living with HIV are typically several years younger than HIV positive men because women tend to have sex with slightly older men.

Using prospectively collected data from people of all ages in a population of roughly 69,000 living in rural northeast South Africa (Kahn et al., 2007), I look for these signature

effects of HIV on sex-age-specific mortality through time. Because one should expect to see important changes in child mortality and a well-defined 'hump' in the age pattern of mortality for adults, I summarize the age pattern of mortality using the eight-parameter Heligman-Pollard mortality model (Heligman and Pollard, 1980). This model decomposes the age pattern of mortality into three pieces, each with a small number of parameters to control it. There are three parameters to describe child mortality (A, B & C), three to describe a very flexible *accident* hump typically occurring in young adulthood (D, E & F) and finally two parameters to describe mortality at older ages (G & H). Recognizing that the 'accident' hump could just as easily represent the much larger bulge in the age pattern of mortality for adults dying of AIDS, I reinterpret this as the 'AIDS' hump and apply the model with this interpretation in the South African setting.

The Heligman-Pollard model is fit to age-specific mortality schedules describing different periods of the HIV epidemic to yield a time series of values for the parameters A-H. Because each parameter controls a specific component of the shape of the age pattern of mortality, the parameters have specific interpretations. As a result, the time series of parameter values describes changes in the age pattern of mortality in a succinct and informative way. For example, different values of D, E & F describe changes in the *location*, *level* and *spread* of the adult mortality hump as HIV becomes more prevalent.

The Heligman-Pollard model is a natural choice in this application, and consequently it is surprising that it has not already been used to describe HIV-related mortality.[1] The likely explanation is that it is hard to fit the Heligman-Pollard model using standard procedures such as minimization of squared errors (Rogers, 1986; Dellaportas, 2001). A solution is to apply the Bayesian melding method (Poole and Raftery, 2000) that has the added advantage of properly quantifying uncertainty in the estimated parameters and the mortality age patterns output by the model. The resulting probability distributions of the parameters and mortality age patterns are used to argue that significant changes have occurred to the age pattern of mortality in ways that are consistent with the effects of

[1]The HP is not alone in this respect. Despite numerous parametric configurations to capture the shape of human mortality (Gage and Mode, 1993; Wood et al., 2001; Carriere, 1992; Siler, 1979), these models are rarely if ever applied to developing world settings especially Africa. One could argue this lack of use is expected given the dearth of data on age-specific mortality in these regions and the fact that many of these models were generated with developed world data and intended to replicate developed world mortality patterns.

HIV.

This chapter is organized as follows. I begin with the background and significance of the work and a review of the age-specific impact of HIV on mortality, followed by a detailed description of the data, Heligman-Pollard model, and Bayesian estimation procedure. I next summarize the findings and detail future work in this area. Finally, the appendices contain the raw data and additional, detailed results.

2.2 Background & Significance

The results from this chapter contribute to a growing literature on the link between HIV prevalence and mortality. I identify age-specific increases in all-cause mortality that follow increases in HIV prevalence with a predictable lag. These results confirm our general understanding of how HIV epidemics work and corroborate and add to the specific findings of other researchers, briefly reviewed below.

2.2.1 Age-specific Effect of HIV on Mortality

The HIV virus attacks a person's immune system, gradually wearing it down until it cannot no longer control infections, or even itself. The result is a long period of infection and gradually worsening illness ending in death. Because of the long lag time between infection and death, the effect of HIV on mortality is observed several years after infection. In the absence of antiretroviral treatment, the time between infection and death for children is between five and ten years (Marston et al., 2005) and for adults about ten years (Jaffar et al., 2004; Morgan et al., 2002). The age pattern of the effect of HIV on mortality is determined by the age pattern of transmission and whether or not there is widespread use of antiretroviral drug therapy (Highly Active Anti-Retroviral Therapy, HAART) in the population. The common modes of transmission in high prevalence populations are heterosexual sexual intercourse and mother-to-child transmission at birth or during breastfeeding. Consequently, infections occur at or shortly after birth and at ages when people are most sexually active. Those ages vary from population to population but generally span the late teens through older adulthood, with peaks sometime in the twenties or thirties. Moreover, individual variation in sexual activity effectively protects a fraction of the population and creates a gradient of risk in the remaining fraction. The

net result in a mature HIV epidemic in a population without widespread treatment is that the bulk of the at-risk portion of the population is infected soon after becoming sexually active, and as a consequence the effect of HIV on mortality is relatively concentrated at the youngest possible ages, about ten years after the average age at infection. Finally, because women typically pair with slightly older men, the average age at infection for women is usually several years younger than for men, and hence the mortality effect of HIV is slightly younger for women compared to men. In general this leads to an age-profile of HIV-related mortality that affects infants and young children, women roughly aged 25-50, and men roughly aged 30-60. Although there is a lot of variation in this general pattern, depending on the specifics of HIV transmission and whether or not HAART is available, this general sex-age-pattern of HIV mortality is commonly observed in populations with high prevalence (Porter and Zaba, 2004).

2.2.2 HIV & Mortality in sub-Saharan Africa

In sub-Saharan Africa the HIV epidemic affects all segments of the population stretching beyond high-risk groups and into the most rural communities (Poit et al., 2001). Rising HIV prevalence over the past 15 to 20 years has led to increases in child and adult mortality and dramatic decreases in life expectancy (Tollman et al., 1999; Hosegood et al., 2004). Our understanding of the role of HIV in shaping mortality in sub-Saharan Africa is supported by information from a wide variety of data, including both direct and indirect estimates of mortality. Indirect evidence on adult mortality from national and regional statistics, including censuses and sample surveys, provide an understanding of trends in mortality at national or regional levels (Blacker, 2004), while community-based and cohort studies offer more direct evidence on the role of HIV/AIDS by comparing the mortality experience of infected and non-infected individuals (Zaba et al., 2004, 2007).

Most of Africa does not have a functioning vital registration system that covers enough of the population to produce indicators that are either representative or accurate.[2] Instead, population censuses and sample surveys have become the main source of represen-

[2]Data on mortality in developing regions can often be of meager quality due to poor vital registration systems (Porter and Zaba, 2004). In addition, people with HIV/AIDS often die of some other more immediate cause, which is then recorded as the primary cause of death resulting in some underreporting of HIV/AIDS-related mortality. These issues extend to South Africa as well (Botha and Bradshaw, 1985; Tollman et al., 1999).

tative information on demographic trends. These data reveal consistent, large increases in adult mortality (measured as ${}_{45}q_{15}$ or the probability a 15 year old will die before reaching age 60) since the early 1990s for many countries in sub-Saharan Africa (Blacker, 2004). Based on an analysis of sibling histories in Demographic and Health Surveys, Timæus and Jasseh (2004) report that by the year 2000 an average 15 year old person living in sub-Saharan Africa faced a probability of dying before age 60 of between 0.3 and 0.6, up from 0.1 to 0.3 during the 1980s.

An important limitation of these data is that they do not describe the causes of death. Nonetheless, they still provide strong evidence that HIV is largely responsible for the recent upswing in mortality rates. The sibling history study reported by Timæus and Jasseh (2004) reveals a sharp increase in adult mortality after countries develop a generalized (at least 1% of the population infected[3]) epidemic. Likewise, the United Nations (2004) report that in countries where adult mortality was either declining or stabilizing, HIV prevalence was 5% or lower, while countries with increasing mortality had prevalence between 7 and 33%. National-level data also suggest a strong effect of HIV on age-specific mortality. Blacker (2004) notes that an examination of the age pattern of mortality increases is useful in assessing the impact of HIV on adult mortality; specifically rapidly rising adult mortality that peaks at younger ages for women compared to men. Timæus and Jasseh (2004) note that the excess mortality they observe in the DHS data is concentrated at ages 25-39 for women and 30-44 for men.

Complementing the ability of census and survey data to describe macro-level trends in mortality, community-based studies that collect data on the HIV serostatus of participants can compare the mortality of infected and uninfected individuals. The age profile of the effect of HIV on mortality is determined by the stage of the epidemic, the age profile of incidence in the past and the average time between infection and death. Because community studies have the ability to track the survival times of infected individuals, they are one of the only sources of data able to provide average times between infection and death. Studies like this that have prospectively monitored cohorts of HIV positive

[3]A stationary population with an expectation of life of ten years, similar to the HIV-positive population, has a crude death rate of $1/10 = 0.1$ or 100 deaths per 1,000 people. A population containing 1% HIV-positive people will therefore add about 1 additional death per 1,000 people to the crude death rate.

people estimate survival times in the range 9-11 years for individuals infected in their 20s and shorter survival times for people infected at older ages (Porter and Zaba, 2004; Todd et al., 2007). Direct evidence from community-based studies provides the best understanding of the effect of HIV on the level and age pattern of adult mortality (see for example: Porter and Zaba, 2004; Groenewald et al., 2005; Adjuik et al., 2006; Nyirenda et al., 2007; Smith et al., 2007; Zaba et al., 2007; Marston et al., 2007). Data from health and demographic surveillance system (HDSS) field sites in Africa and Asia – all prospective community-level sites – identify seven age patterns of mortality, two of which likely reflect a substantial effect of HIV. Those two patterns have significant humps in the age pattern of mortality between ages 20 to 55 for males and 20 to 45 for females (Clark, 2002).

South Africa is experiencing one of the most rapidly progressing HIV epidemics in the world (Hosegood et al., 2004; Karim and Karim, 1999; Department of Health, Republic of South Africa, 2003). While there was virtually no HIV mortality at the beginning of the 1990s, by 2000 Dorrington et al. (2001) identify significant HIV-related mortality. Hosegood et al. (2004) report that AIDS was the most important single cause of death driving the steep increase in mortality rates in South Africa. As in other parts of sub-Saharan Africa, HIV prevalence among South African women is younger than among men, and the risk of dying from AIDS peaks at younger ages for women (25-39) compared to men (30-44) (Tollman et al., 1999; Hosegood et al., 2004; Groenewald et al., 2005).

HIV is also affecting child mortality in sub-Saharan Africa, and again there are few data to describe this effect. The data that do exist come largely from small hospital-based studies and population-based projects that estimate HIV-related child mortality rates (Zaba et al., 2004) using a variety of methods. Although the effect of HIV on under-five mortality varies broadly by region, the impact appears to be substantial in southern Africa where the worst affected countries are; in these populations HIV may be causing up to half of all child deaths (Newell et al., 2004). Some 90% of pediatric HIV infections occur in sub-Saharan Africa and because many HIV-infected children die before their fifth birthday, childhood mortality overall can be greatly elevated by HIV (Dabis and Ekpini, 2002; Foster and Williamson, 2000; De Cock et al., 2000). Given the high rates of transmission from mother to child, one could expect that increases in HIV prevalence of

adult women will increase child mortality. HIV may also indirectly affect child mortality through maternal HIV status because children of HIV infected mothers are more likely to die than those of uninfected mothers (Newell et al., 2004; Zaba et al., 2004).

2.3 Methods & Data

2.3.1 The Heligman-Pollard Mortality Model

I use the Heligman-Pollard mortality model (Heligman and Pollard, 1980), shown in equation 2.1, to assess mortality at all ages over a 14-year period. This model expresses the probability of dying as a function of age using a three-part expression that covers the entire age range.

$$q_x = A^{(x+B)^C} + D \times e^{-E(ln(x)-ln(F))^2} + \frac{GH^x}{1+GH^x} \tag{2.1}$$

where x indexes age[4]*, q_x is the probability of death at age x, and $A, B, C, \ldots, H$ (see Table 2.1) are eight variable parameters that govern the shape of the mortality curve*[5]

Compared to non-parametric alternatives, the advantages of parametric models like the Heligman-Pollard are: 1) the ability to produce smooth curves, 2) the ability to interpolate and to extrapolate, 3) the possibility of formal manipulation, 4) parsimony, 5) interpretable parameters whose values can be compared easily and 6) the ability to easily capture and reflect trends and mediate comparisons (Dellaportas, 2001; Debón et al., 2005). These last two advantages, along with the ability of parametric models to succinctly summarize large amounts of data, are especially useful for the present study. The representation of mortality trends via parametric methods facilitates a time-series approach whereby the shape and intensity of age-specific mortality curves can be summarized and compared over time (Debón et al., 2005; Congdon, 1993; Rogers and Gard, 1991; Hartmann, 1987).

The Heligman-Pollard model has been used to document changes in mortality in a variety of contexts (see for example: Heligman and Pollard (1980); Rogers (1986); Forfar

[4]The HP cannot be expressed at age 0. For the youngest age one can use a very small number like 0.00001 or the formula described in Rogers and Gard (1991, p. 80). For simplicity I take the former approach and evaluate the model at ages $x \in \{0.00001, 1, 5, 10, \ldots, 100\}$.

[5]The HP model expresses the probability of dying at some age, q_x. I interpret these values to be the probability of death in the interval x to $x+n$, where n is the length of the interval.

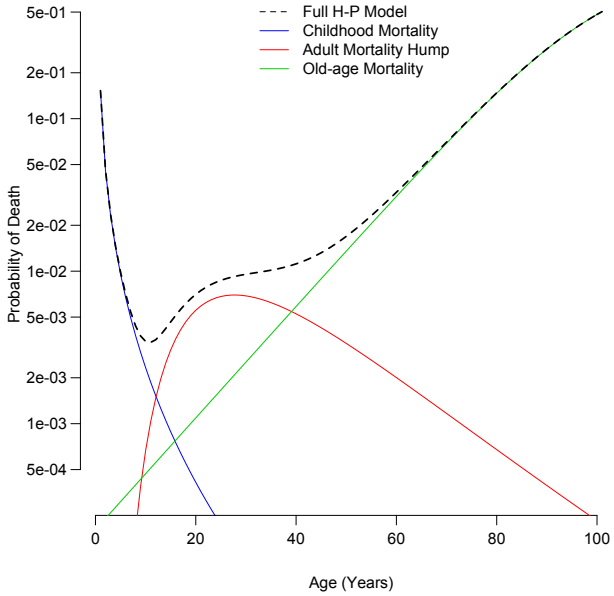


Figure 2.1: **Decomposition of the Heligman-Pollard Model into its three constituent age compnents. The blue, red, and green lines plot the childhood, adulthood, and old-age components of the model respectively, while the dashed black line shows the complete all-age model.**

and Smith (1987); Hartmann (1987); Congdon (1993); Dellaportas (2001); Debón et al. (2005)). In their original paper Heligman and Pollard (1980) describe Australian mortality over the 20th century, and Rogers and Gard (1991) document declining infant and young adult mortality over the 20th century in the U.S. There are few examples of this type of use for the Heligman-Pollard mortality model in the developing world, and there does not appear to be a previously published example of the Heligman-Pollard mortality model being used to characterize mortality schedules affected by HIV over time.

Although originally conceived to capture the slight rise in mortality from accidents occurring in the late teens and early adult years, the so-called 'accident' hump contained in the Heligman-Pollard mortality model is flexible enough to represent the much larger bulges in adult mortality brought about by AIDS deaths. The ability of the Heligman-Pollard to reflect levels of mortality at all ages and in particular to characterize the level and shape of adult mortality make it well suited to assess the increasing intensity of adult mortality over the 1990s and 2000s in sub-Saharan Africa. Moreover, because the model parameters have straightforward demographic interpretations (i.e. describing not only the

level but also the shape and location of adult mortality) one can easily link the parameter changes to lagged HIV prevalence.

The eight parameters of the Heligman-Pollard model control three separate additive components corresponding to three age ranges of the mortality schedule (child mortality, young adult mortality and late life mortality), and each parameter has a demographic interpretation (Heligman and Pollard, 1980; McNown and Rogers, 1989; Rogers and Gard, 1991). Figure 2.1 plots each of the components and the sum of all three, i.e. the final curve.[6] Table 2.1 provides interpretations for the parameters, and the eight panels of Figure 2.2 illustrate the effect of each of the parameters individually. Each panel of this figure plots the modeled mortality schedule with variation in a single parameter while holding the others constant.

Table 2.1: **Heligman-Pollard Model Parameters**

Parameter	Description
A	Level of child mortality, approximately the probability of dying between ages 1 and 2, ${}_1q_1$
B	Difference between age 0 and 1 mortality probabilities
C	Decline in mortality during childhood
D	Level (or height) of mortality hump
E	Inversely related to the width of the mortality hump
F	Location of the mortality hump on the age axis
G	Late life mortality, intercept of Gompertz curve at age 0
H	Late life mortality, slope of of Gompertz curve

The first three parameters, A, B & C, control the first component and describe early child mortality. A roughly approximates mortality at age one and can be taken as a measure of the intensity or level of child mortality (McNown and Rogers, 1989; Rogers and Gard, 1991; Hartmann, 1987). B is the age displacement variable (Rogers and Gard, 1991) and reflects the difference between mortality at ages 1 and 0. As the value of B increases, ${}_1q_0$ or the probability a new born will die before their first birthday, decreases

[6]Figures 2.1 and 2.2 are plotted using $A = 0.06008$, $B = 0.31087$, $C = 0.34431$, $D = 0.00698$, $E = 1.98569$, $F = 26.7107$, $G = 0.00022$, $H = 1.08800$ which correspond to the nearly 100 ${}_nq_x$ values in the Brass standard (McNown and Rogers, 1989).

below 0.5 and begins to approach ${}_1q_1$, the probability a one year old dies before their second birthday (Rogers and Gard, 1991). Finally, C indicates how quickly mortality decreases during childhood and into the young adult years. Parameters A, B & C all have domain (0, 1). Declines in A are consistent with decreasing child mortality (Hartmann, 1987). Because the potential direct impact of pediatric AIDS deaths and the indirect impact of adult AIDS mortality on child mortality, the level of child mortality should increase from period to period leading to increases in A for both sexes as time goes on and the HIV epidemic grows.

The second component of the model is designed to reproduce the 'accident' hump in many male mortality schedules, and possibly maternal mortality in female mortality schedules (Heligman and Pollard, 1980; Hartmann, 1987; McNown and Rogers, 1989; Rogers and Gard, 1991). When it reflects accidents, the hump peaks in the (early) twenties. I reinterpret the hump to reflect AIDS mortality at young and slightly older adult ages, so in this study the peak is likely to be older and taller. Parameter D controls the level or intensity of young adult mortality, E is inversely related to the spread of the hump and F locates its position along the age axis (Heligman and Pollard, 1980; McNown and Rogers, 1989; Rogers and Gard, 1991). Parameters D and E have domains (0, 1) and $(0, \infty)$ respectively while the domain of F is less clear. I restrict F to the domain (15, 55) because a large number of AIDS deaths at ages older than 55 is unlikely. With the hump reflecting AIDS mortality, one should expect increasing levels of adult mortality for both sexes as time progresses, i.e. increases in D. As the epidemic matures we should also see a broadening and aging of the adult mortality hump, i.e. decreases in E and increases in F. Note that if D is sufficiently small the other two parameters controlling this component of the curve have no influence. Inspecting equation 2.1 and the panel for parameter D in Figure 2.2, one can see that the second component of the model is composed of the product of D and the effects of E and F, so a small D essentially negates the effect of the spread and location parameters.

The last two parameters control the late life mortality section of the curve and describe the steep increase in mortality at those ages.[7] Parameter G measures the base level of

[7]Kostaki (1992) formulated a nine-parameter version of the HP with a Makeham function as the old-age term instead of Gompertz. Since the focus of this paper is adult mortality and to a somewhat lesser extent, child mortality, I choose the more parsimonious formula without the added complexity

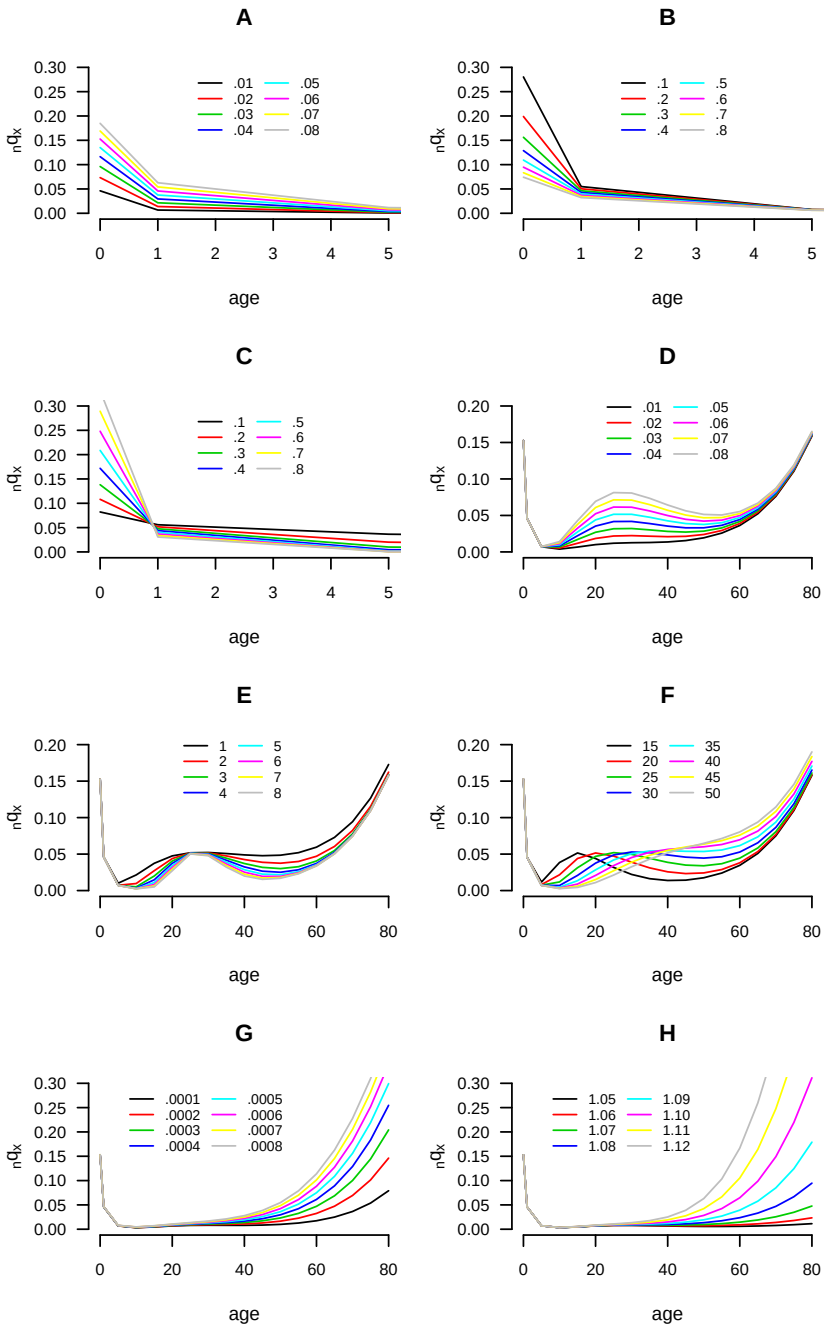


Figure 2.2: **Variation Plots for the Heligman-Pollard Mortality Model.** ${}_nq_x$ refers to the probability of death in the interval x to $x+n$. Each panel plots the age-specific probabilities of death when varying a single parameter in the HP model while holding all others constant. For the plots of parameter E and F, D is set to .04 to give sufficient intensity to the hump.

mortality at these ages ($x = 0$) and H defines the rate of increase (Rogers and Gard, 1991). Parameters G and H are the intercept and slope of the Gompertz curve respectively and take domains (0, 1) and (0, ∞) respectively.

The aim is to identify values for the eight parameters that produce a curve that matches a set of mortality measures at various ages. The set of eight parameter values constitutes a compact, interpretable description of mortality, and the curve the parameters represent is a smoothed version of the empirical data.

2.3.2 Data

For this study, I use data describing age-specific mortality and HIV prevalence. Mortality data come from the Agincourt HDSS in South Africa (Kahn et al., 2007). The site is located in a rural area of Mpumalanga Province in the northeast of South Africa where HIV prevalence is a little less than 30% (Department of Health, Republic of South Africa, 2006). Using time-sex-age-specific counts of deaths and person years for the period 1994-2007, I compute time-sex-age-specific probabilities of dying (${}_nq_x$) using standard life table methods. The estimation procedure used for this study requires the denominator in the probability of dying (l_x or the number of people left alive at age x) to have approximately the same scale as the empirical data.[8] Thus the l_x column of the life table needs to be rescaled. I do this using MS Excel's 'solver' utility to find a value for l_0 that yields the same total number of person years, T_0, in the life table as there are in the observed data.

Because the annual data contain few deaths and are rather 'noisy,' I aggregate across time. Periods are chosen to contain years with similar mortality rates and to capture as much variability in the age pattern of mortality as possible. The four periods I use are: 1994-1997 ($l_0^{male} = 1,917$, $l_0^{female} = 1,873$), 1998-2001 ($l_0^{male} = 2,100$, $l_0^{female} = 2,049$), 2002-2004 ($l_0^{male} = 1,896$, $l_0^{female} = 1,822$) and 2005-2007 ($l_0^{male} = 1,974$, $l_0^{female} = 1,893$). Figure 2.3 displays the aggregated and rescaled mortality schedules for males and females. Appendix A lists the data as used in the present analysis and Appendix F details the process for aggregating into these periods.

---

of a ninth parameter.

[8] I do not know the actual number of people left alive at each age, l_x, but it needs to be approximated closely or the uncertainty estimates will be inaccurate. For example, if I set the radix of the life table to 100,000 as is common, the resulting confidence intervals are far too narrow because there are far fewer actual people at the radix.

Data describing the trend in HIV prevalence in Mpumalanga Province are taken from reports published by the South African Department of Health (1995; 1996; 1997; 1998; 2003; 2006). For the most part these estimates are made using anonymous HIV test results from pregnant women who attend antenatal clinics. See section 2.3.4 for further description on this data source.

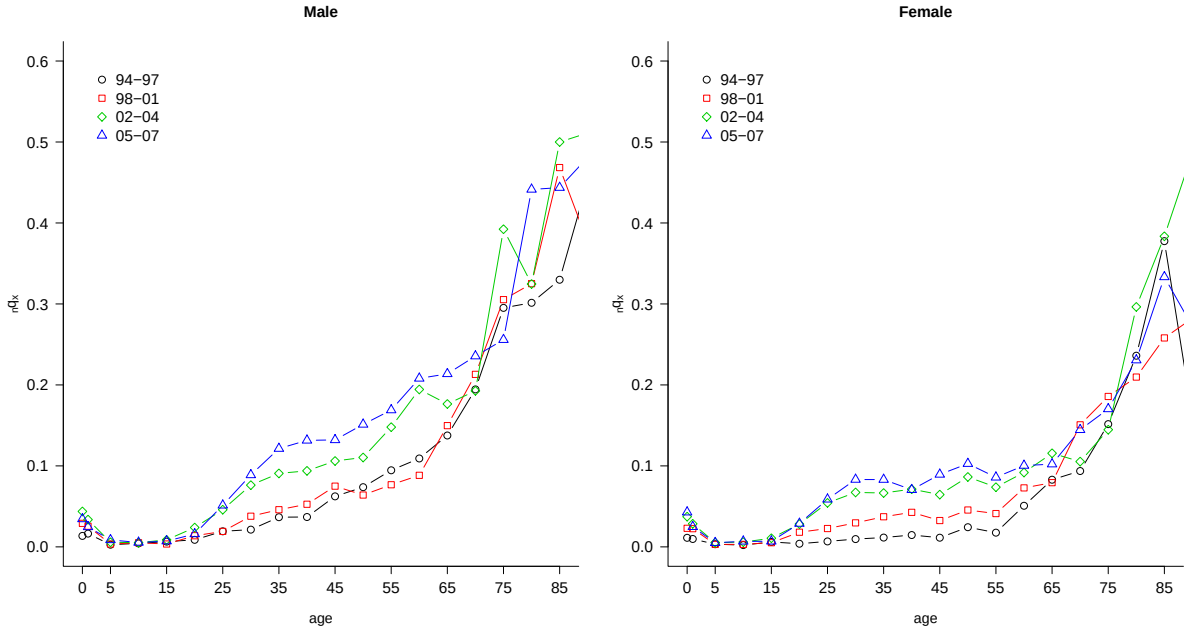


Figure 2.3: **The Age Pattern of the Probability of Death for Agincourt HDSS for four aggregated periods: 1994-1997, 1998-2001, 2002-2004, and 2005-2007.** ${}_nq_x$ refers to the probability of death in the age interval x to $x+n$

2.3.3 Bayesian Melding Estimation

Bayesian melding searches the joint parameter space for sets of parameter values that are most likely given the data and what is known beforehand about the typical values for the parameters and the age patterns of mortality being modeled. Instead of identifying just the one most likely set of parameter values, it finds the most likely region of the parameter space and returns the sets of parameter values in that region. Each set of parameter values is associated with a probability corresponding to how well it reflects the data. These probabilities are used to construct joint distributions of the parameter values, which allow one to draw inferences about and characterize uncertainty around the

parameter estimates themselves and the mortality patterns they represent.[9]

In the Bayesian framework parameters are treated as random variables. Prior beliefs about the parameters are quantified in the form of a joint probability density $p(\theta)$, where θ is a vector of parameters for which we will make inference. The data $\mathbf{y}$ are brought in by specifying a likelihood $\mathcal{L}(\mathbf{y}|\theta)$, which is the probability of the observed data for a given value of the parameter vector. I use the binomial likelihood, which is a natural choice given that the observed ${}_nq_x$ are probabilities (see appendix G for specification of the likelihood). Bayesian melding (Poole and Raftery, 2000) is designed for problems in which a deterministic model – such as the Heligman-Pollard mortality model – is used in the likelihood function. Let M represent the model which transforms a set of parameter inputs θ into a set of model outputs $\phi = M(\theta)$. The prior density for the model inputs $p(\theta)$ and a likelihood for the outputs and the data $\mathcal{L}(M(\theta))$ are combined to produce the posterior distribution for the model inputs. Using Bayes' Theorem and the marginal density of the data $p(\mathbf{y})$, one can update prior beliefs to obtain the posterior distribution[10]

$$p(\theta|\mathbf{y}) \propto \mathcal{L}(\mathbf{y}|M(\theta))p(\theta) \tag{2.2}$$

which is used to make inference for θ.

For this analysis of the Heligman-Pollald mortality model, I specify independent uniform priors that are intended to be fairly uninformative, placing most of the influence with the observed data. With these priors, the aim is to exert as little influence as possible on the parameter outputs while keeping the parameter output values in plausible ranges. The prior distributions are

$$\begin{array}{lll} A \sim U[0, 0.25] & D \sim U[0, 0.25] & G \sim U[0, 0.01] \\ B \sim U[0, 1] & E \sim U[0, 20] & H \sim U[1, 1.5] \\ C \sim U[0, 1] & F \sim U[15, 55]\,. & \end{array}$$

---

[9]The following discussion of Bayesian melding adapted from (Clark et al., 2010). For a detailed discussion of Bayesian Melding see Poole and Raftery (2000).

[10]Equation 2.2 arises from the fact that the marginal density of the data, $p(\mathbf{y})$, does not depend on θ, so the posterior distribution only needs to be known up to a constant and is thus proportional to the product of the likelihood and the prior.

Inference is performed by sampling from $p(\theta|\mathbf{y})$ and summarizing the resulting posterior sample. One can evaluate the model for each set of inputs in the posterior sample to obtain a posterior sample of the model outputs $p(\phi|\mathbf{y})$. Note that the posterior sample reflects the distribution of model outputs, and thus the quantiles of the posterior sample can be used to make probabilistic statements about the values of the model outputs (q_x).

In this analysis, I use the Incremental Mixture Importance Sampling algorithm (see appendix H for a discussion of IMIS) to sample 400 sets of parameter values in the final resample from the posterior distribution, which can then be used to calculate 400 separate ${}_nq_x$ schedules, each covering the whole age range. The result is an approximation of the posterior distribution of ${}_nq_x$ values, i.e. a distribution of the probabilities of dying, *not* the number of deaths, which is what was used in the likelihood and therefore what is desired. To produce a predicted distribution of the number of deaths that I can compare to the data, I simulate death counts from the posterior distribution of ${}_nq_x$ values. This is accomplished using the binomial distribution with the observed number of people at risk of dying and an ${}_nq_x$ value sampled at random from the posterior distribution of ${}_nq_x$. The number of deaths is then divided by the number of people at risk of dying to produce a predicted ${}_nq_x$ value. This procedure is repeated many times to yield a distribution of predicted ${}_nq_x$ values. Figure 2.4 plots the ${}_nq_x$ values simulated from the posterior output distribution for the first (flat hump) and the last (more intense and concentrated hump) periods for males. The gray clouds are the simulated ${}_nq_x$ values. This figure shows how the simulated, predicted ${}_nq_x$ values cluster around the observed age-specific death probabilities and that this model can closely fit ${}_nq_x$ schedules at the beginning and peak of the HIV epidemic. Appendix D presents the prior and posterior densities for each parameter by sex and year.

In the current application, the Bayesian melding approach has important advantages compared to common alternative estimation procedures. The Heligman-Pollard model is usually fit using least squares methods, but because the number of parameters (8) is large in comparison to the number of data points in a typical age pattern of mortality (20), the model is on the verge of being over-parameterized (Bebbington et al., 2007; Dellaportas, 2001; Congdon, 1993). Dellaportas (2001) use a different Bayesian approach and report that the over-parameterization issue is usually resolved by using informative priors

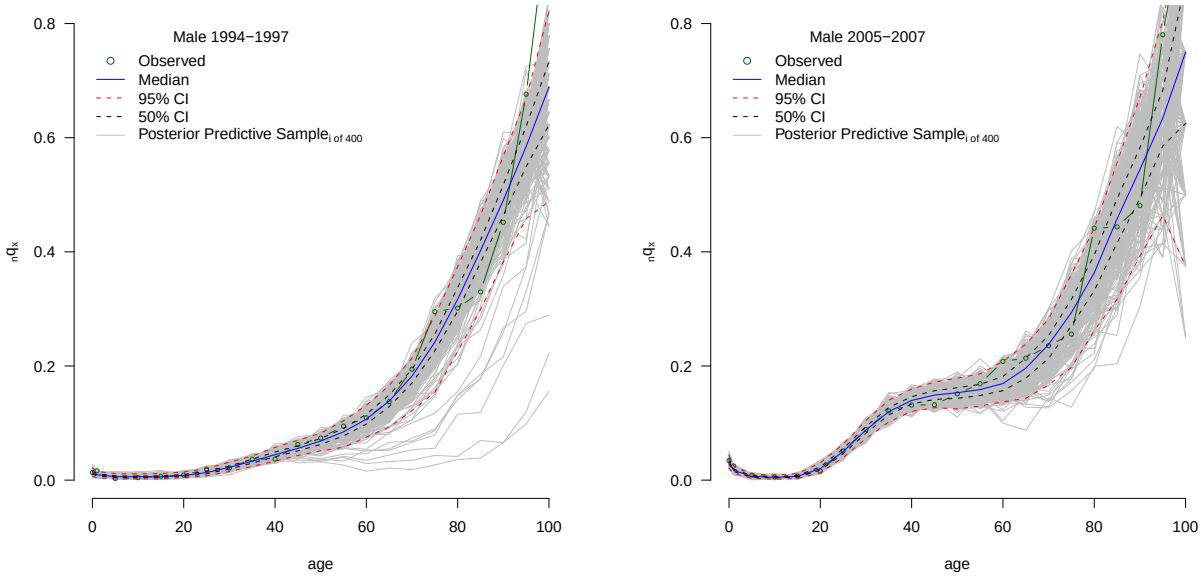


Figure 2.4: **Predictive Distributions of Male Age-Specific Probabilities of Dying for 'Flat' Hump ('94-97) and 'Intense' Hump ('05-07) Periods.** ${}_nq_x$ refers to the probability of death in the interval x to $x+n$. Predictive distributions result from fitting the periods with the Heligman-Pollard model estimated with Bayesian Melding.

(Congdon, 1993). Even with the relatively uninformative priors used here, the Bayesian approach solves this problem. One way to assess over-parameterization is to examine the correlation between parameters. If there is a strong correlation between parameters, over-parameterization may be an issue. Pairs plots of the posterior parameter samples for each sex/year combination are presented in Appendix C. The parameters are essentially uncorrelated except G and H, and the correlation of G and H is expected since they act together as a product. Another difficulty in estimating the Heligman-Pollard model occurs when adult mortality is somewhat flat and the estimated hump parameters can take values beyond a plausible range (Rogers and Gard, 1991; Hartmann, 1987; Heligman and Pollard, 1980). In this case, the level parameter D may be large and the location parameter F on order 100, essentially creating a very broad, flat hump whose upswing merges into the natural rise in mortality at older ages. Fixing one or more parameters and estimating the others tends to resolve this problem (Congdon, 1993; Rogers and Gard, 1991; Hartmann, 1987). Instead, I let all the parameters vary simultaneously while restricting the range on parameter F to between 15 and 55, which keeps the hump parameters in plausible ranges and prevents the hump from merging into old-age mortality.

2.3.4 Relationship between Parameters and HIV Prevalence over Time

To investigate the temporal relationship between trends in the values of the Heligman-Pollard parameters and HIV prevalence, for each sex I plot the median value of the parameter distributions for parameters A, D, E and F by the HIV prevalence in Mpumalanga Province with a five-year lag (figure 2.11). I expect increases in HIV prevalence to be followed by increases in both child and adult mortality, so there should be a positive relationship between both child and adult mortality and the level parameters A and D. Additionally, the spread and location parameters E and F should reveal changes in the general age structure of adult mortality affected by HIV. Table 2.4 contains the HIV prevalence estimates for the relevant five-year lagged periods.

2.3.5 Calculating Uncertainty around Life Table Columns

An advantage of the Bayesian approach is the ability to estimate uncertainty bounds around both the model parameters and the *outputs*. Following a similar method advanced by Lynch and Brown (2005), I exploit this property to generate uncertainty bounds around the other columns of a standard life table, in particular around the life expectancy, e_x, schedules.

For each period and sex, the predictive ${}_nq_x$ distributions, shown in Appendix B, contain 400 individual ${}_nq_x$ schedules from which I calculate a 'distribution' of 400 individual life tables. I present the median e_x, life expectancy at at x, distributions in figure 2.12 in Appendix C and the median e_0 and e_{10} with measures of uncertainty summarized from their distributions in Tables 2.4 and 2.6. I use these to make statistically valid statements about changes in the expectation of life.

2.4 Results

Figure 2.7 displays the period-specific fits, each period in a different color. For each period the light-colored dots in the background are the observed data, the solid lines are the median fitted curves and the dashed lines define the 95% credible intervals around the median, which are not required to be symmetric. The radical increases in age-specific mortality are obvious: massive increases for infants and very young children and for adults aged roughly 15-65. The increases are not uniform with age and peak for men in the early

40s and for women slightly younger. The corresponding decreases in life expectancy are listed in Table 2.4: men lost fifteen years of life expectancy, and women lost fourteen. *The substantive objective of this work is to decompose these changes into their constituent parts and relate those parts to changes in HIV prevalence.*

Table 2.5 presents the median parameter estimates and 95% credible intervals. Figure 2.8 plots the estimated values of parameters A, D, E and F over the four periods in this study. The vertical dashed lines in this figure represent the 95% credible interval around each point estimate.

2.4.1 Goodness of Fit

Before discussing the parameters themselves, I assess goodness of fit by calculating the percentage of data points contained within posterior credible intervals (CI) of varying lengths. Tables 2.2 and 2.3 contain the percentage of data points contained in the 50, 80, 90, 95, and 99 percent posterior credible intervals for three age ranges, 15-60 (adult mortality), 0-60, and 0-95 for males and females respectively. For each CI we should expect about the same percentage of data points to be contained within the interval. For virtually all years and both sexes, all of the data points in the adult ages are contained within the 99% CI with roughly 90 percent of data points in this age range in the 90 and 95% CIs. Adding in the childhood ages (rows labeled '0-60') and oldest ages (rows labeled '0-95') we see the model again does a good job of fitting these data for both sexes and all periods.

Numerous other parametric forms to capture the shape of human mortality have been proposed (Gage and Mode, 1993; Wood et al., 2001), so in addition to showing how well the Heligman-Pollard model fits these data, one might also wish to know how well the HP model fits compared to other well-known parametric functions. I fit the four periods for both sexes with two other parametric models (using maximum likelihood), Gompertz-Makeham (Gompertz, 1825; Makeham, 1860) and Siler (Siler, 1979, 1983), and show those fits with the median HP fit. Figures 2.5 and 2.6 plot these fits for all four periods. These plots make clear that the HP is able to best replicate all-age mortality in the presence of high HIV prevalence. During the first two periods when mortality is relatively flat in the adult years, all models do relatively well but when the AIDS hump begins to grow, the

Table 2.2: The percentage of data points (observed age-specific probabilities of death) covered in 50, 80, 90, 95 and 99% posterior credible intervals after fitting the Heligman-Pollard model to four aggregated periods of data, 1994-1997, 1998-2001, 2002-2004, 2005-2007, from Agincourt HDSS. Male

Ages	50% CI	80% CI	90% CI	95% CI	99% CI
Male (94-97)					
15-60	50	80	90	90	100
0-60	50	71	79	86	100
0-95	48	67	71	86	100
Male (98-01)					
15-60	50	90	90	90	100
0-60	50	86	86	86	100
0-95	43	76	81	81	90
Male (02-04)					
15-60	60	90	90	90	100
0-60	57	86	86	86	93
0-95	48	71	81	81	86
Male (05-07)					
15-60	50	80	90	90	100
0-60	50	79	86	86	100
0-95	43	76	86	90	100

other models do not have have the flexibility that the HP does with its three parameters governing the adult hump.

2.4.2 Child Mortality

The increase in child mortality during the period when the HIV epidemic grows is reflected by increases in parameter A, which approximates the probability of dying between ages one and two, ${}_1q_1$. The upper-left panel of Figure 2.8 displays a clearly upward trend in the value of parameter A for both male and female children. A quick visual comparison of the credible ranges (dashed vertical line in figure 2.8) reveals unambiguous increases in parameter A during the first two intervals and a possible decrease during the last interval.

Table 2.3: The percentage of data points (observed age-specific probabilities of death) covered in 50, 80, 90, 95 and 99% posterior credible intervals after fitting the Heligman-Pollard model to four aggregated periods of data, 1994-1997, 1998-2001, 2002-2004, 2005-2007, from Agincourt HDSS. Female

Ages	50% CI	80% CI	90% CI	95% CI	99% CI
Female (94-97)					
15-60	40	90	90	90	90
0-60	36	93	93	93	93
0-95	33	76	81	81	81
Female (98-01)					
15-60	50	70	100	100	100
0-60	43	57	93	100	100
0-95	38	48	76	81	90
Female (02-04)					
15-60	60	90	90	90	100
0-60	50	86	93	93	100
0-95	48	81	86	90	100
Female (05-07)					
15-60	50	60	80	90	100
0-60	50	64	86	93	100
0-95	52	76	90	95	100

2.4.3 Adult Mortality

Parameters D, E, F, G and H control adult mortality. For this discussion, I focus on D, E and F because they control the hump while G and H define underlying adult mortality. The plots in figure 2.8 reveal striking changes in the hump parameters (upper right and lower two panels). The level parameter D increases significantly over all three intervals for both sexes except during the first interval for males with little to no overlap in the credible intervals from one period to the next. Figures 2.9 and 2.10 plot the period-sex-specific hump component of the model in isolation without the childhood or old-age components. Figure 2.9 plots the hump for each sex by period and clearly reveals the growth in the hump from period to period. Figure 2.10 displays the hump curves with the male and female humps in a single plot for each period, so that the sexes can be compared easily. This figure confirms that although men appear to have slightly higher humps in all

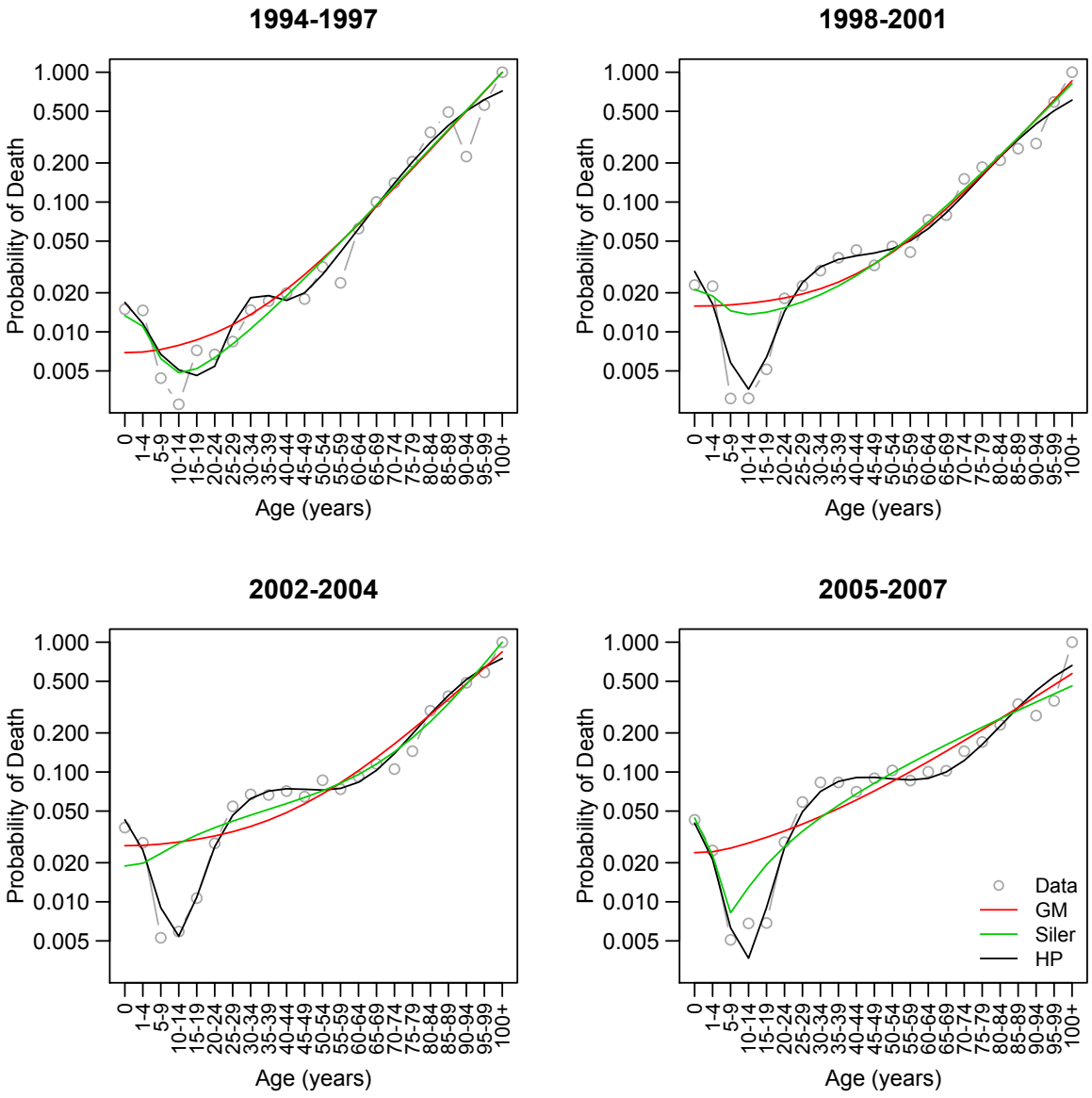


Figure 2.5: **Median HP fit and maximum likelihood fit for Gompertz-Makeham and Siler models to the four periods of Agincourt data, female**

periods, the differences between the sexes are small and insignificant except for the last period when women have a substantially lower hump. These movements in parameter D relate directly to the growth of the mortality bulge observed for adults, and the highly regular and significant changes (i.e. non-overlapping credible intervals) demonstrate how large and important the increases in adult mortality have been.

Trends in the width parameter E (inversely related to the width of the hump) indicate that over time the hump has become significantly wider for women and somewhat wider

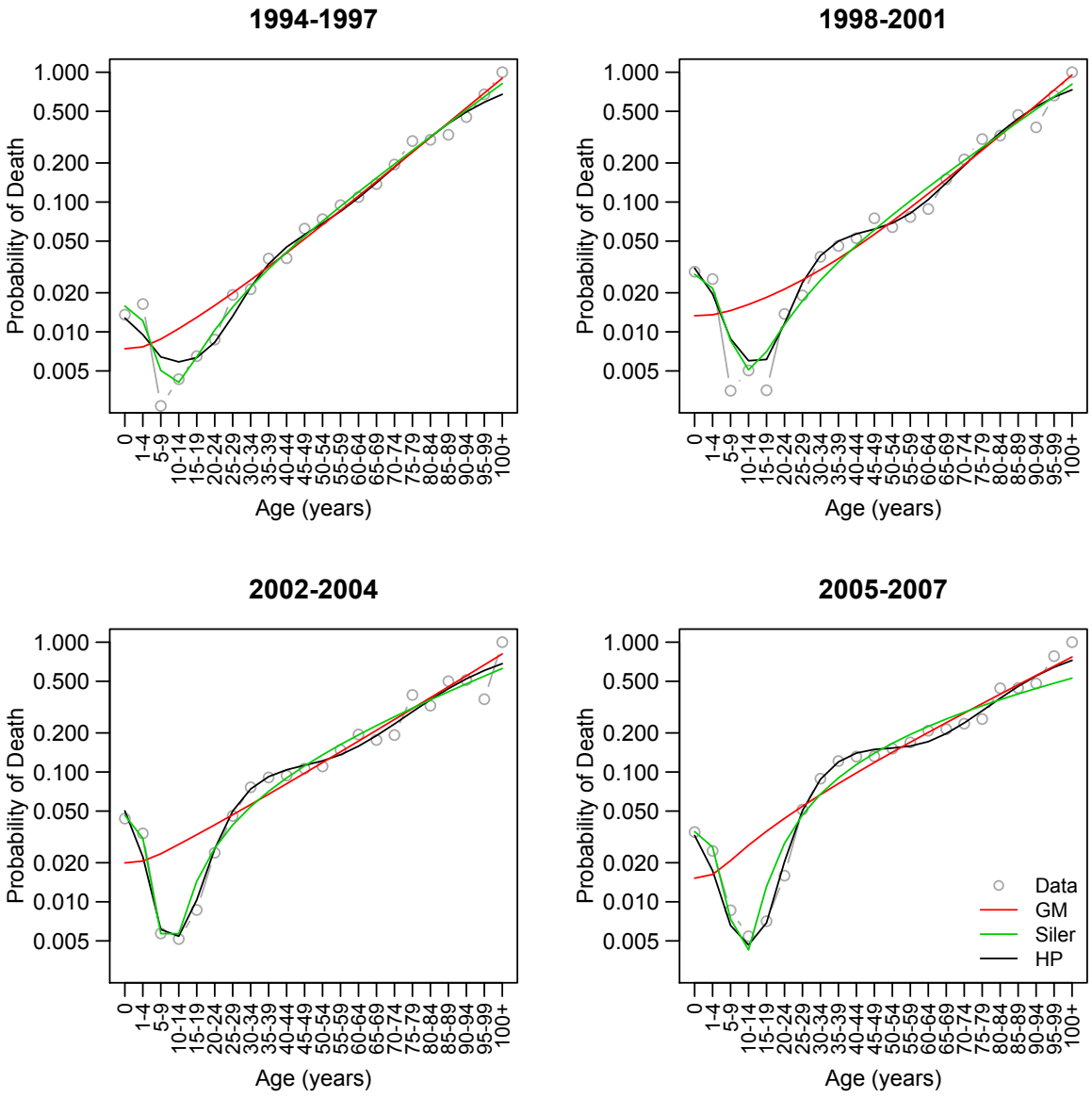


Figure 2.6: **Median HP fit and maximum likelihood fit for Gompertz-Makeham and Siler models to the four periods of Agincourt data, male**

for men, with both converging to a fairly wide configuration and staying that way. The exceptionally high value of the width parameter for women in first period is picking up a slight but concentrated increase in mortality in the late 20s to early 30s, possibly reflecting the earliest and youngest HIV incidence for women. Although the credible intervals overlap for all periods after the initial period, both sexes experience a slight broadening of the hump as time progresses.

Changes in the location parameter F (location of the hump on the age axis) reveal

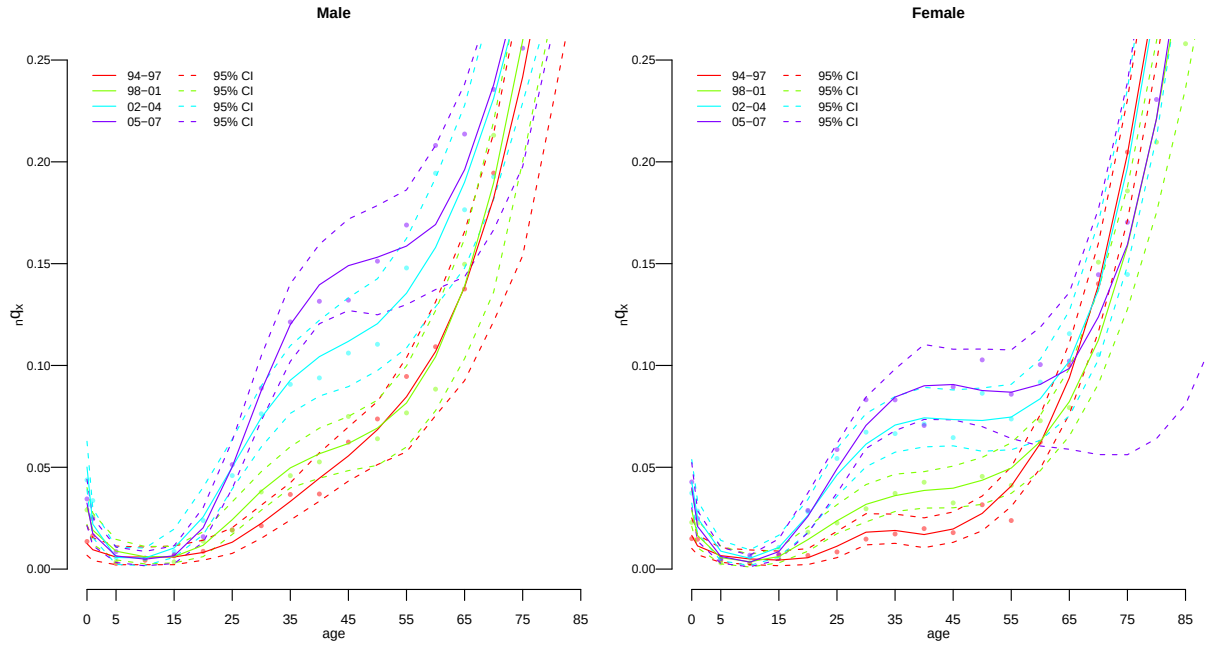


Figure 2.7: **Age-specific Probability of Dying by Period for Agincourt HDSS as estimated by the Heligman-Pollard model.** Median fitted curves (solid lines) with 95% credible intervals (dashed lines). Observed ${}_nq_x$ values in light-colored dots. ${}_nq_x$ refers to the probability of death in the interval x to $x+n$.

that the hump has steadily become older for women and became younger and then older again for men. Figures 2.8 and 2.10 suggest that the hump is typically slightly older for males; the male humps usually peak at slightly older ages regardless of the height or spread of the hump. This is especially pronounced during the first period when the mortality effects of HIV are first being observed. Recall that women tend to be infected with HIV at younger ages compared to men and so have high mortality at younger ages. The location parameter accurately reflects this previously observed pattern with the male median estimate of F consistently outpacing the female estimate. Women experience increasing mortality during their 30s (reflecting a female incidence profile that starts infecting a significant number of women in their early 20s) while for men the hump grows from the late 30s to the mid 40s. The unusually large value of the location parameter in the first period for men probably results from a small increase in mortality around the late 40s to early 50s. Because this is the first period, it is unlikely that this small hump is a result of HIV-related deaths. Rather, HIV probably begins to affect the male mortality schedule between the first and second periods, at which time the location and

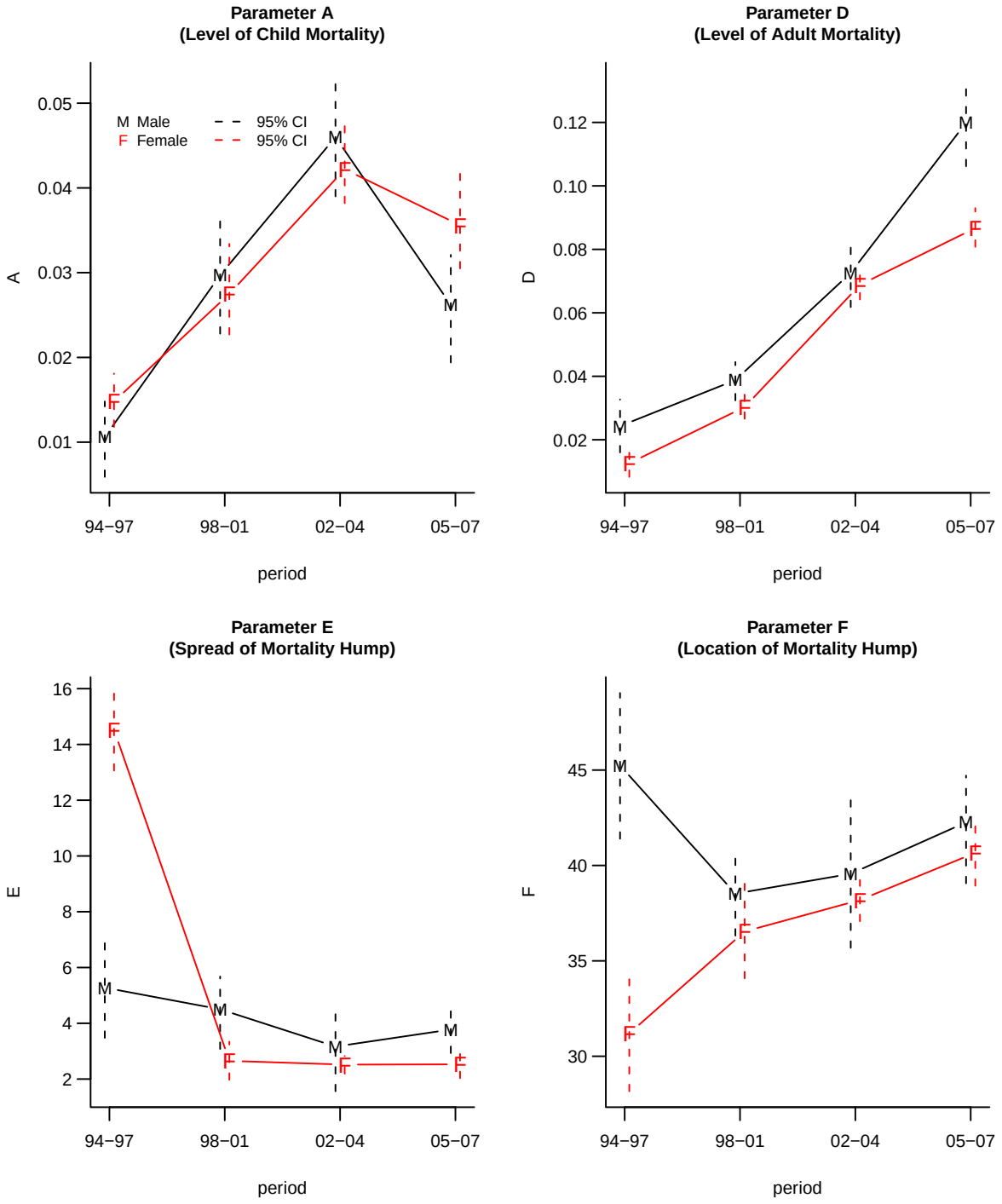


Figure 2.8: **Estimated Values for Parameters A, D, E, F of the Heligman-Pollard Mode by Sex and Time for Agincourt HDSS 1994-1997, 1998-2001, 2002-2004, 2005-2007.**

Table 2.4: **Life Expectancy for Agincourt and HIV Prevalence for Mpumalanga Province**

Period	Male e_0[a]	ΔMale[b]	Female e_0[a]	ΔFemale[b]	Prev.[c]	ΔP[d]	Prev. Period
1994-1997	67.3	—	73.5		1.3	—	1990-1992
	(65.2 – 72.0)		(72.0 – 74.9)				
1998-2001	63.0	-4.3	69.7	-3.8	11.6	10.4	1993-1996
	(61.3 – 65.4)		(67.8 – 71.8)				
2002-2004	53.5	-9.5	59.7	-10.0	26.6	15.0	1997-1999
	(51.6 – 55.4)		(58.1 – 61.9)				
2005-2007	52.2	-1.4	59.4	-0.4	29.2	2.5	2000-2002
	(50.6 – 54.4)		(56.4 – 64.8)				
Total Change:		**-15.2**		**-14.2**		**27.9**	

[a] e_0 in years. 95% credible intervals for e_0 in parentheses.

[b] Δ(Male, Female) is change in median e_0 from previous period.

[c] HIV prevalence is for the surrounding province (Mpumalanga) for all ages in period roughly five years before period in which e_0 is estimated. Source for prevalence estimates: Department of Health, Republic of South Africa 1995; 1996; 1997; 1998; 2003; 2006.

[d] ΔP is change in HIV prevalence from previous period.

level parameters begin to pick up increasing male mortality in the late 30s, consistent with previously observed HIV mortality patterns. Change in the hump for men is seen best in figure 2.9; clearly between the first and second periods the male hump increases slightly in intensity and shifts back to the expected age range.

Comparing humps from the last two periods with those from the first two, two things are clear: 1) the hump becomes taller reflecting the increases in the level parameter, D, and 2) the humps broaden as the width parameter, E, decreases. The widening of the humps is asymmetric with the bulk of the changes happening at older ages (35+) where the hump appears to flatten out. These changes reflect the spreading of increased adult mortality into the older middle adult ages, 35-45, likely reflecting a widening of the age profile of incidence resulting in deaths associated with HIV at increasingly older ages.

Table 2.5: **Heligman-Pollard Parameter Estimates (95% CI)**

	1994-1997	1998-2001	2002-2004	2005-2007
		MALE		
A	0.0106	0.0297	0.0460	0.0262
	(0.0059 - 0.0148)	(0.0228 - 0.0365)	(0.0390 - 0.0529)	(0.0194 - 0.0321)
B	0.7829	0.9724	0.9652	0.7754
	(0.6673 - 0.8979)	(0.9248 - 0.9980)	(0.8928 - 0.9981)	(0.6930 - 0.8609)
C	0.0855	0.1791	0.3526	0.2083
	(0.0320 - 0.1501)	(0.1149 - 0.2336)	(0.2801 - 0.4191)	(0.1573 - 0.2589)
D	0.0243	0.0388	0.0726	0.1199
	(0.0160 - 0.0327)	(0.0324 - 0.0445)	(0.0618 - 0.0836)	(0.1062 - 0.1338)
E	5.2779	4.5021	3.1653	3.7761
	(3.4764 - 7.1116)	(3.0783 - 5.6768)	(1.5618 - 4.6151)	(2.9309 - 4.6871)
F	45.2412	38.5225	39.5636	42.2610
	(41.4002 - 49.0394)	(36.3215 - 40.5921)	(35.6854 - 43.7599)	(39.0647 - 44.7113)
G	0.0010	0.0006	0.0021	0.0009
	(0.0003 - 0.0017)	(0.0003 - 0.0009)	(0.0010 - 0.0034)	(0.0004 - 0.0015)
H	1.0798	1.0877	1.0717	1.0827
	(1.0725 - 1.0872)	(1.0822 - 1.0939)	(1.0630 - 1.0805)	(1.0772 - 1.0887)
		FEMALE		
A	0.0148	0.0275	0.0422	0.0356
	(0.0118 - 0.0181)	(0.0227 - 0.0334)	(0.0382 - 0.0473)	(0.0305 - 0.0417)
B	0.7852	0.9305	0.9792	0.8563
	(0.7047 - 0.8732)	(0.8589 - 0.9942)	(0.9520 - 0.9985)	(0.8085 - 0.9027)
C	0.1048	0.2120	0.2245	0.2373
	(0.0697 - 0.1409)	(0.1621 - 0.2568)	(0.2065 - 0.2447)	(0.2064 - 0.2669)
D	0.0126	0.0303	0.0688	0.0867
	(0.0083 - 0.0167)	(0.0266 - 0.0348)	(0.0642 - 0.0733)	(0.0808 - 0.0930)
E	14.5269	2.6527	2.5174	2.5266
	(13.0611 - 15.8447)	(1.9731 - 3.3408)	(2.1914 - 2.8271)	(2.0314 - 3.1009)
F	31.1959	36.5381	38.1891	40.6617
	(28.1650 - 34.0361)	(34.0823 - 39.2770)	(37.0759 - 39.2378)	(38.9361 - 42.2265)
G	0.0003	0.0003	0.0001	0.0001
	(0.0002 - 0.0004)	(0.0002 - 0.0004)	(0.0000 - 0.0001)	(0.0000 - 0.0002)
H	1.0958	1.0900	1.1110	1.1092
	(1.0904 - 1.1019)	(1.0824 - 1.0972)	(1.1049 - 1.1170)	(1.0976 - 1.1174)

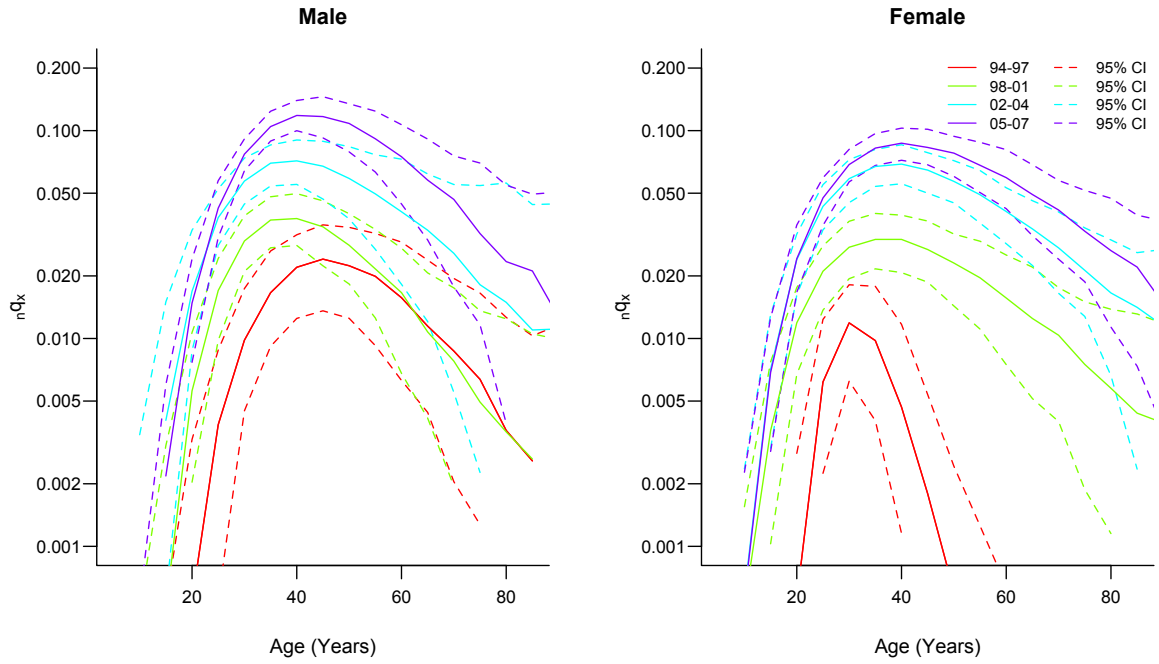


Figure 2.9: **Hump Component for All Periods and Both Sexes, Agincourt HDSS 1994-1997, 1998-2001, 2002-2004, 2005-2007.** The solid lines are the median of the posterior distribution with the dashed lined plotting the 95% posterior credible interval.

2.4.4 HIV Prevalence

Figure 2.11 displays the median parameter values for A, D, E and F along with the corresponding HIV prevalence five years earlier from Table 2.4. The period labels near each point on the plots correspond the the mortality estimation periods associated with the parameter values. The HIV prevalence values come from periods five years before the label on each point. The gray dashed line in each panel is a simple OLS regression line to help the eye assess the linearity of the relationship between the parameter values and HIV prevalence.

For all periods and both sexes, the level parameters A and D increase with HIV prevalence, except for A during the final period. During the final interval the overall level of child mortality is reduced, perhaps reflecting the success of programs to limit mother-to-child transmission (for a brief review see Doherty et al., 2005). The largest increase in level is in the second interval (between the second and third periods) for women coming just after the largest increase in prevalence; the largest increase for men occurs in the last interval after ten years at high prevalence.

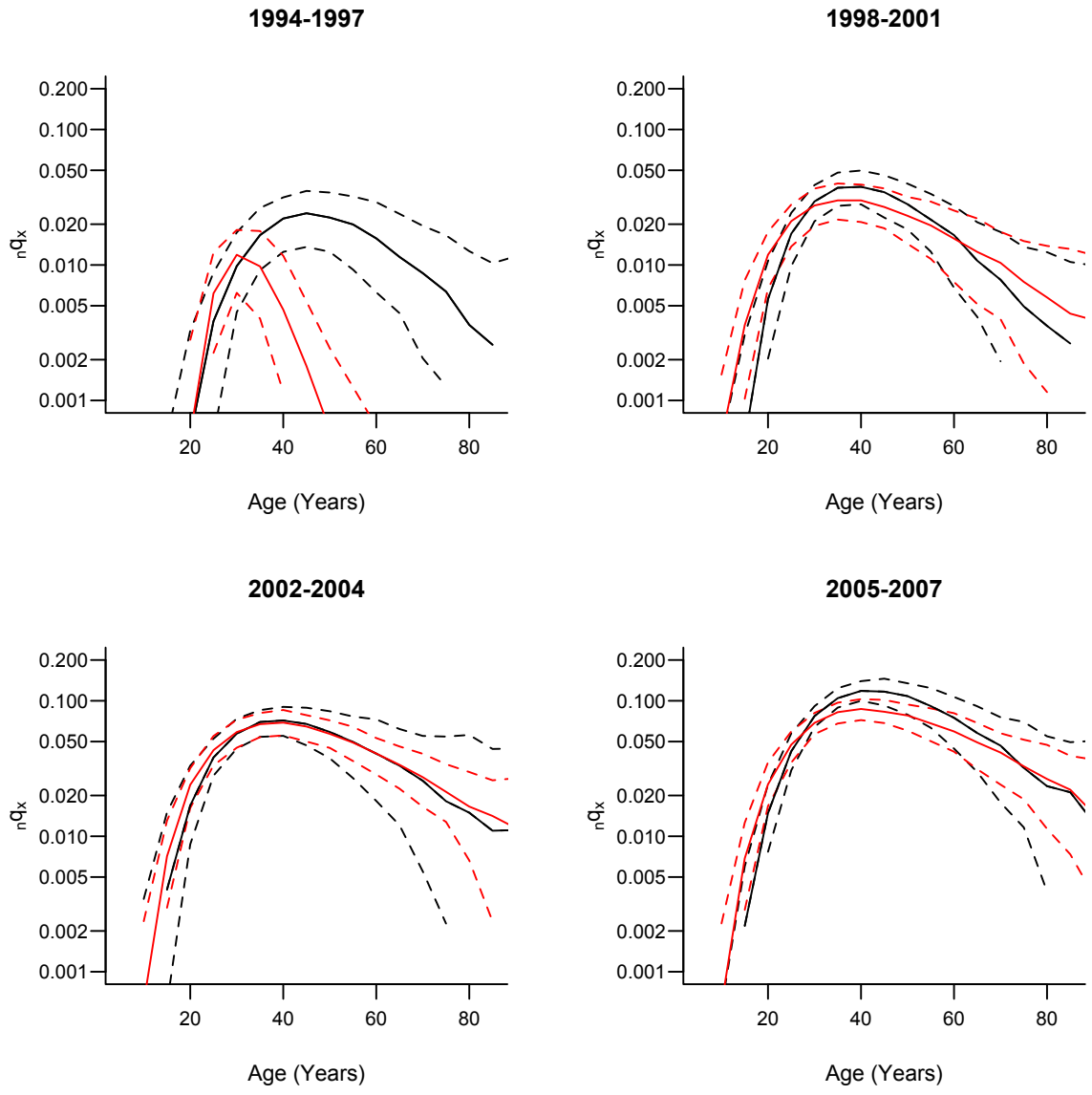


Figure 2.10: **Hump Component for Each Sex for Each Period, Agincourt HDSS 1994-1997, 1998-2001, 2002-2004, 2005-2007.** The solid lines are the median of the posterior distribution with the dashed lined plotting the 95% posterior credible interval.

For parameters E (spread or width of hump) and F (location of hump) the relationship with earlier prevalence is less neat but still regular and in-line with expectations. For both sexes there is a downward slope to the relationship between E (width of the hump) and earlier prevalence indicating that as prevalence increases the hump becomes gradually wider; this is especially the case for women whose hump starts off much narrower than men's.

For men the location of the hump moves from ~ 45 to ~ 38 and then back up to ~ 42. In contrast, women have experienced a consistent 'aging' of the hump moving from ~ 31 to ~ 40. This indicates the age *locus* of the mortality impact of HIV for men has hovered at a little below age 40 while for women it has moved steadily to older ages and is now at about the same age as men.

2.4.5 Life Expectancy

Table 2.6 reports expectation of life at birth and age ten for the four periods for both sexes, and Figure 2.12 plots the median expectation of life schedule with 95% credible intervals. There was a precipitous decline in life expectancy over the entire period with the largest drops occurring between the second and third periods for both sexes. At the beginning of observation during 1994-1997 a male infant could expect to live 67.3 years, and a female infant 73.5. By the end of the final period, a male infant expected only 52.2 years and a female 59.4. Men lost 15.2 years of life expectancy, and women lost 14.2. Judging from the credible intervals, change over the whole period is very significant, with the bulk of the significant drop between periods two and three.

The expectation of life at age ten is not affected by child mortality and consequently describes adult mortality more clearly. Table 2.6 shows clearly that the majority of the changes in the expectation of life at birth result from changes in adult mortality; the changes in e_{10} closely track, and in the most recent period, exceed the changes in e_0. The timing of this trend in the expectation of life is consistent with the rise in HIV prevalence. Given a mean survival time of 9-11 years between infection with HIV and death in the absence of access to antiretroviral therapy, one would expect (and indeed is observed) rapid increases in mortality at about the turn of century, reflecting increases in HIV prevalence during the 1990s, see Table 2.4.

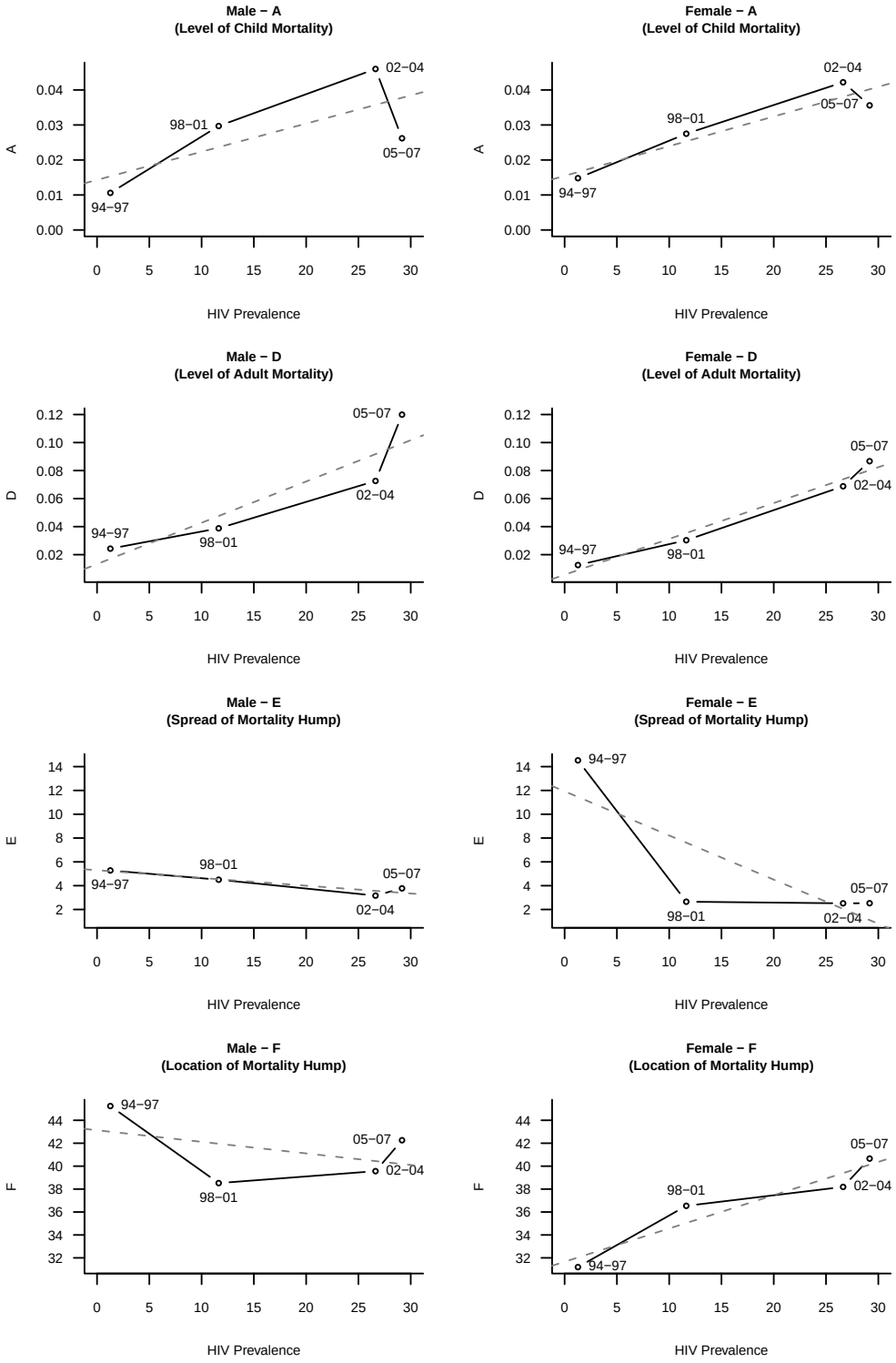


Figure 2.11: **Selected Parameter Values by 5-year Lagged HIV Prevalence** (OLS regression line in dashed grey).

Table 2.6: **Life Expectancy at Birth and Age 10**

Period	e_0[a]	Δe_0[b]	e_{10}[a]	Δe_{10}[b]
		MALE		
1994-1997	67.3	—	59.2	—
	(65.2 – 72.0)		(57.2 – 64.2)	
1998-2001	63.0	-4.3	56.8	-2.5
	(61.3 – 65.4)		(55.0 – 59.4)	
2002-2004	53.5	-9.5	47.9	-8.9
	(51.6 – 55.4)		(46.1 – 49.8)	
2005-2007	52.2	-1.4	45.1	-2.7
	(50.6 – 54.4)		(43.8 – 47.4)	
Total Change:		**-15.2**		**-13.1**
		FEMALE		
1994-1997	73.5	—	66.0	—
	(72.0 – 74.9)		(64.7 – 67.3)	
1998-2001	69.7	-3.8	63.4	-2.6
	(67.8 – 71.8)		(61.8 – 65.6)	
2002-2004	59.7	-10.0	54.5	-9.0
	(58.1 – 61.9)		(53.1 – 56.5)	
2005-2007	59.4	-0.4	53.4	-1.0
	(56.4 – 64.8)		(50.4 – 59.0)	
Total Change:		**-14.2**		**-12.6**

[a] e_x in years. 95% credible intervals for e_x in parentheses.

[b] Δe_x is change in median e_x from previous period.

2.5 Discussion

2.5.1 Summary

In this chapter I use the parametric Heligman-Pollard mortality model to characterize the age-specific impact of HIV on mortality in the Agincourt study population living in rural northeast South Africa over the period 1994-2007. The eight parameters of the Heligman-Pollard model describe the shape of the age-specific ${}_nq_x$ schedule in three age ranges. I estimate the model using the Bayesian melding method with IMIS to produce

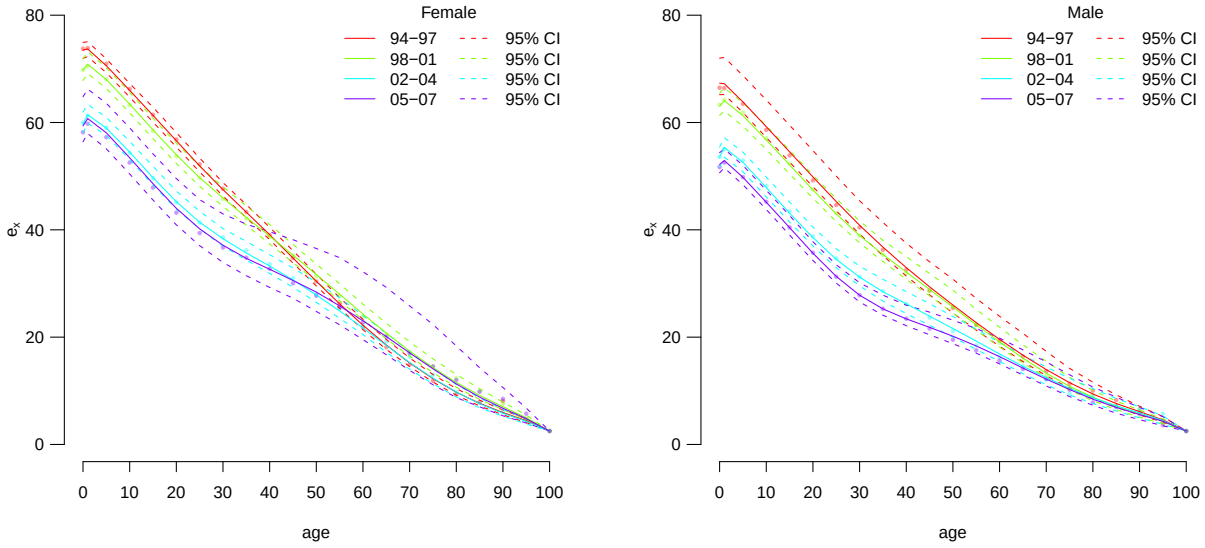


Figure 2.12: e_x **Schedule Predictive Distribution with Median and 95% CI after fitting the Heligman-Pollard Model with Bayesian Melding to data from Agincourt HDSS 1994-1997, 1998-2001, 2002-2004, 2005-2007.** Observed data are represented as lightly colored dots.

robust estimates of the parameter values, the ${}_nq_x$ schedules output by the model and the corresponding life tables in four periods during which the HIV epidemic grew rapidly in this population. The Bayesian framework yields probability distributions of these estimates, and I use these distributions to compare both parameter estimates and ${}_nq_x$ values over time and between the sexes.

The evidence from Agincourt demonstrates the profound effects of HIV on mortality. With a lag of several years after increases in HIV prevalence of 10-15%, the mortality of young children and adults increased dramatically. The corresponding drops in the expectation of life at birth removed about 15 years of life from the average person living in the Agincourt area. All of these changes are highly unlikely to have occurred by chance. Although this is not a new finding, it contributes an unusual nuance–the sex-specific age pattern of increases in mortality following increases in HIV prevalence succinctly summarized with an elegant parametric model.

The eight-parameter Heligman-Pollard model fits HIV-related mortality at Agincourt very well and allows one to summarize the complex age-specific increases in mortality with a small number of estimated parameter values. The parameters that should respond

to HIV-related mortality do indeed reflect the increases in mortality associated with HIV in a temporal sequence consistent with changes in HIV prevalence. The increases in young child mortality are captured with parameter A, and the appearance and growth of the hump in adult mortality are reflected by changes in the 'hump parameters' – D, E and F. The hump level parameter progressively increases, and the hump spread parameter decreases, corresponding to a gradual widening of the hump. Sex differences in these dynamics suggest that the epidemic started and stays slightly older for men, and for women the epidemic started in a very narrow, young age range and gradually expanded to include older ages. The estimated parameters succeed as a parsimonious and informative description of the age-specific changes in mortality as the HIV epidemic grew.

The Bayesian melding estimation procedure allows one to make robust fits of the eight-parameter Heligman-Pollard model of age-specific mortality to period-sex-age-specific mortality rates covering four phases of the growing HIV epidemic in the Agincourt study population. This Bayesian technique also allows one to characterize uncertainty in both the parameters and outputs of the model (the estimated mortality age schedule) without assuming any special properties or relationships among the parameters. Bayesian melding results in a joint posterior distribution of the parameters. Running this posterior through the model yields a posterior distribution of ${}_nq_x$ schedules from which counts of deaths can be simulated and transformed into predicted ${}_nq_x$ schedules, from which predicted life tables can be constructed. This results in a distribution of life tables, or equivalently, a predictive distribution for each column in the life table from which I calculate measures of uncertainty. The resulting estimates and credible intervals allow me to make strong probability-based statements about differences between mortality schedules and components of mortality schedules.

R package Finally, I have released an `R` package, `HPbayes` (Sharrow, 2011) that implements all of the methods described in this paper, and I will continue to develop and improve that package. The package is available as a standard `R` package from the Comprehensive R Archive Network (CRAN) that can be run using the `R` statistical software. The user supplies age-specific counts of deaths and person years, and the package estimates the Heligman-Pollard model for those data and produces a variety of outputs similar to those presented here.

Appendix A: Observed Person Years & Deaths

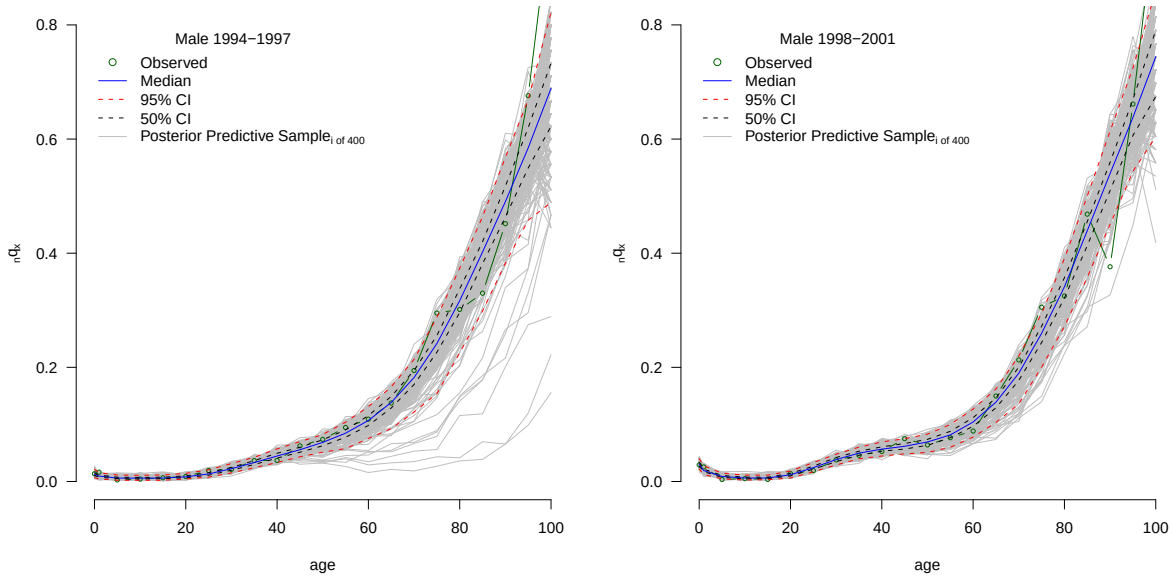
Table 2.7: **Female Person Years & Deaths from Agincourt HDSS during four periods: 1994-1997, 1998-2001, 2002-2004, 2005-2007**

	1994-1997		1998-2001		2002-2004		2005-2007	
Age	PY	Deaths	PY	Deaths	PY	Deaths	PY	Deaths
0	3,771.5	59	3,581.8	83	2,382.8	91	2,567.8	114
1-4	16,100.3	60	14,373.1	82	10,216.0	75	9,555.2	60
5-9	20,370.5	17	19,605.3	12	13,282.0	13	12,666.4	11
10-14	17,551.3	9	18,797.4	12	14,319.4	17	13,365.6	18
15-19	14,853.5	21	15,989.0	17	12,842.0	27	13,536.5	17
20-24	13,002.8	17	13,739.6	50	10,774.3	62	11,542.8	67
25-29	10,749.3	18	11,466.2	51	9,056.7	101	9,354.3	113
30-34	9,163.4	27	9,439.5	56	7,358.1	103	7,436.0	129
35-39	7,026.6	25	8,027.5	61	6,220.5	86	6,172.2	107
40-44	5,831.2	23	6,200.4	54	4,953.4	73	5,418.5	79
45-49	3,973.7	14	5,034.3	34	4,130.0	56	4,153.5	78
50-54	3,248.9	21	3,496.2	33	3,174.8	57	3,526.6	76
55-59	3,140.5	15	3,114.2	26	2,298.6	35	2,577.5	47
60-64	3,042.2	38	2,860.8	44	2,139.9	41	2,039.0	43
65-69	2,834.0	60	2,930.6	48	1,903.7	47	1,894.7	41
70-74	1,742.7	53	2,260.2	73	2,091.9	46	1,725.9	53
75-79	1,200.9	55	1,389.5	56	1,091.5	33	1,682.0	63
80-84	441.0	37	720.9	37	731.4	51	710.4	37
85-89	209.1	28	258.2	16	225.7	22	447.0	38
90-94	99.1	6	99.0	9	109.8	14	110.9	7
95-99	37.3	8	40.1	9	35.0	6	55.6	5
100+	46.0	9	21.4	4	9.2	2	14.9	1

Table 2.8: **Male Person-Years & Deaths from Agincourt HDSS during four periods: 1994-1997, 1998-2001, 2002-2004, 2005-2007**

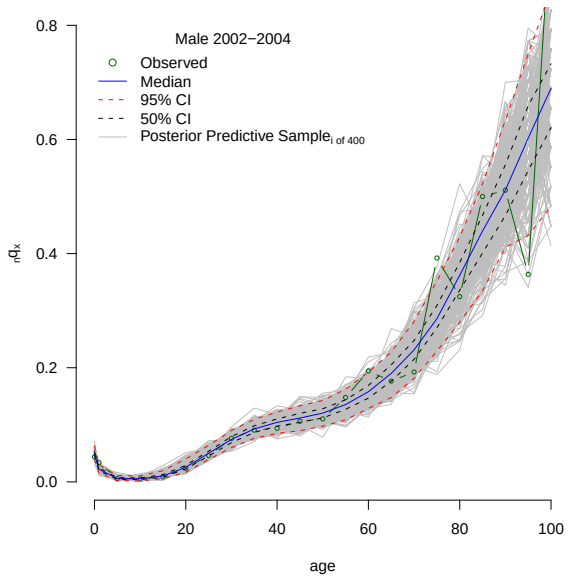
	1994-1997		1998-2001		2002-2004		2005-2007	
Age	PY	Deaths	PY	Deaths	PY	Deaths	PY	Deaths
0	3,795.5	55	3,485.6	103	2,332.6	104	2,587.7	91
1-4	16,183.6	66	14,306.2	91	10,041.9	86	9,321.9	58
5-9	20,452.0	12	19,640.2	12	13,139.5	15	12,326.2	21
10-14	17,469.8	15	18,880.2	17	14,463.3	15	13,113.1	13
15-19	15,055.9	18	16,318.0	13	12,885.9	23	13,803.1	18
20-24	11,798.2	22	13,571.6	36	11,080.0	53	11,514.0	37
25-29	9,543.1	37	10,626.8	41	8,730.0	82	9,418.5	100
30-34	7,803.9	34	8,508.8	65	6,696.8	107	7,005.0	131
35-39	5,863.0	44	6,727.7	63	5,439.4	104	5,505.2	142
40-44	4,833.8	37	5,074.3	56	4,082.4	81	4,488.1	127
45-49	3,810.7	48	4,317.5	67	3,233.4	73	3,244.9	92
50-54	2,698.6	41	3,138.9	41	2,762.8	65	2,759.4	91
55-59	2,461.2	49	2,414.3	38	1,823.0	58	2,102.4	78
60-64	1,617.2	38	2,090.5	38	1,637.4	71	1,389.4	65
65-69	1,529.5	47	1,410.4	45	1,101.4	43	1,350.3	65
70-74	1,190.9	54	1,202.1	57	898.0	39	722.2	39
75-79	848.9	58	888.6	66	556.7	55	674.8	40
80-84	328.6	25	503.9	39	455.1	35	356.1	44
85-89	126.1	12	194.3	25	175.9	23	265.2	31
90-94	45.5	7	51.6	4	60.0	10	69.9	9
95-99	13.5	2	16.8	4	12.4	3	22.2	8
100+	32.7	2	11.0	1	.1	0	3.0	0

Appendix B: Predictive ${}_nq_x$ Schedule Distributions

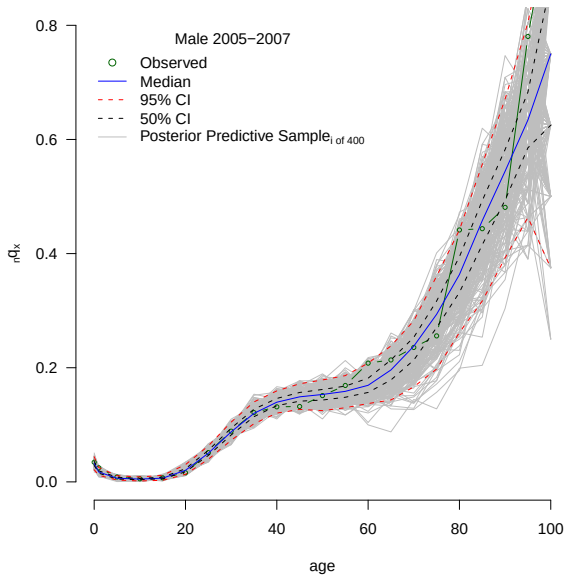


(a) 1994-1997

(b) 1998-2001

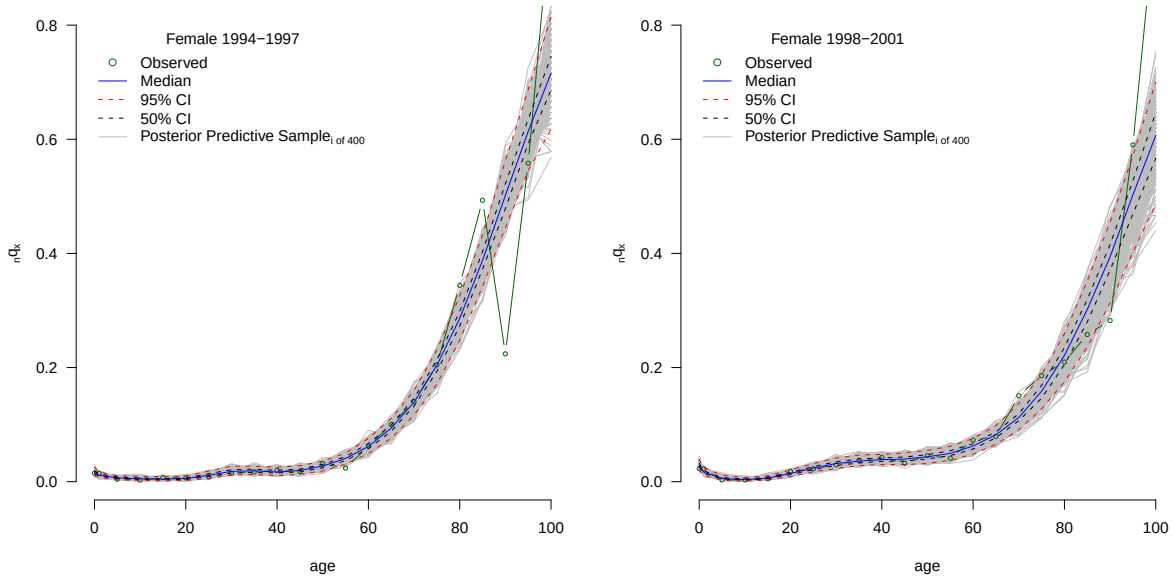


(c) 2002-2004



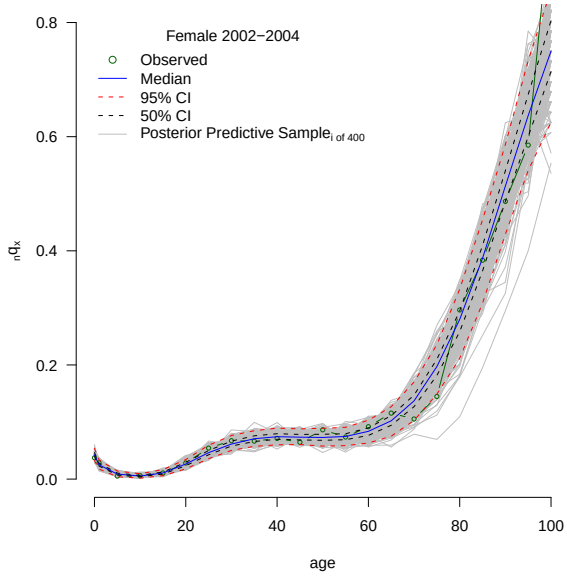
(d) 2005-2007

Figure 2.13: Predictive Distributions of Male Age-Specific Probabilities of Dying for Agincourt HDSS. ${}_nq_x$ refers to the probability of death in the interval x to $x+n$. Predictive distributions result from fitting the periods with the Heligman-Pollard model estimated with Bayesian Melding.

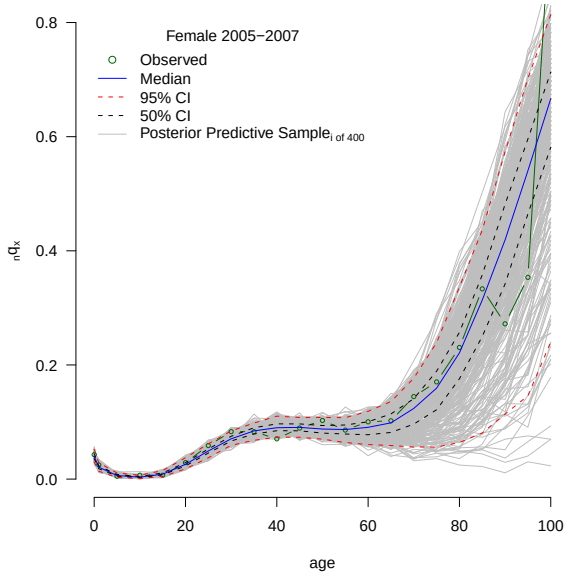


(a) 1994-1997

(b) 1998-2001



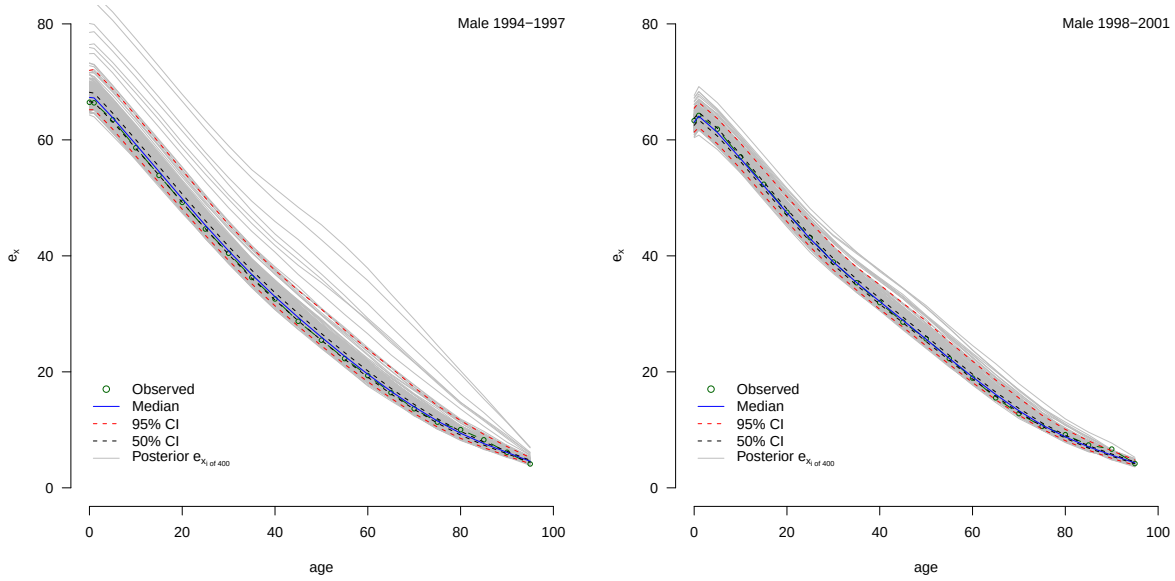
(c) 2002-2004



(d) 2005-2007

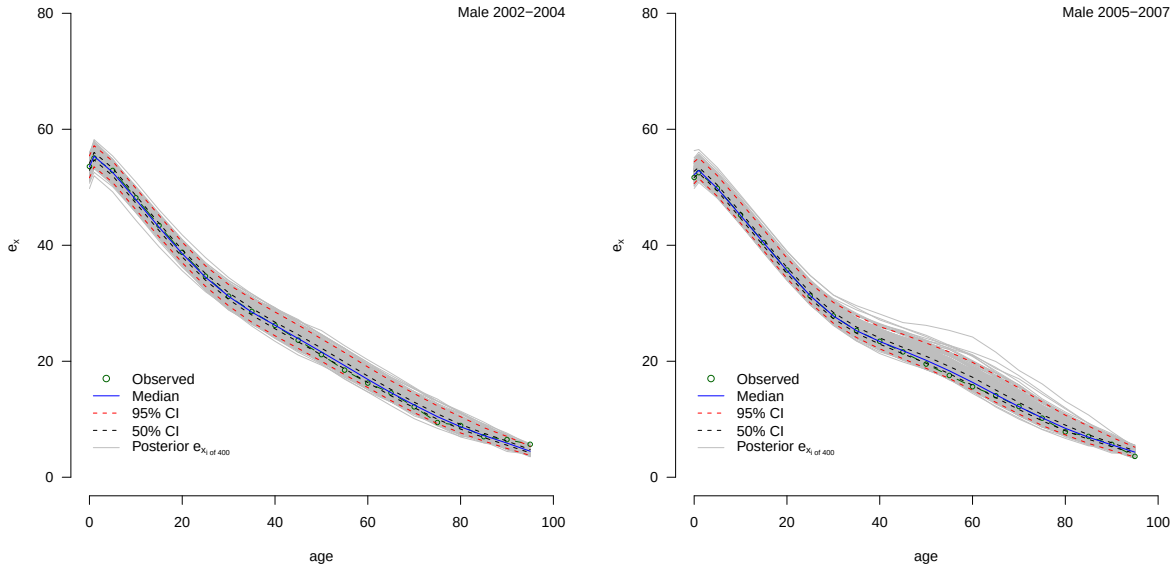
Figure 2.14: Predictive Distributions of Female Age-Specific Probabilities of Dying for Agincourt HDSS. ${}_nq_x$ refers to the probability of death in the interval x to $x+n$. Predictive distributions result from fitting the periods with the Heligman-Pollard model estimated with Bayesian Melding.

Appendix C: Predictive e_x Schedule Distributions



(a) 1994-1997

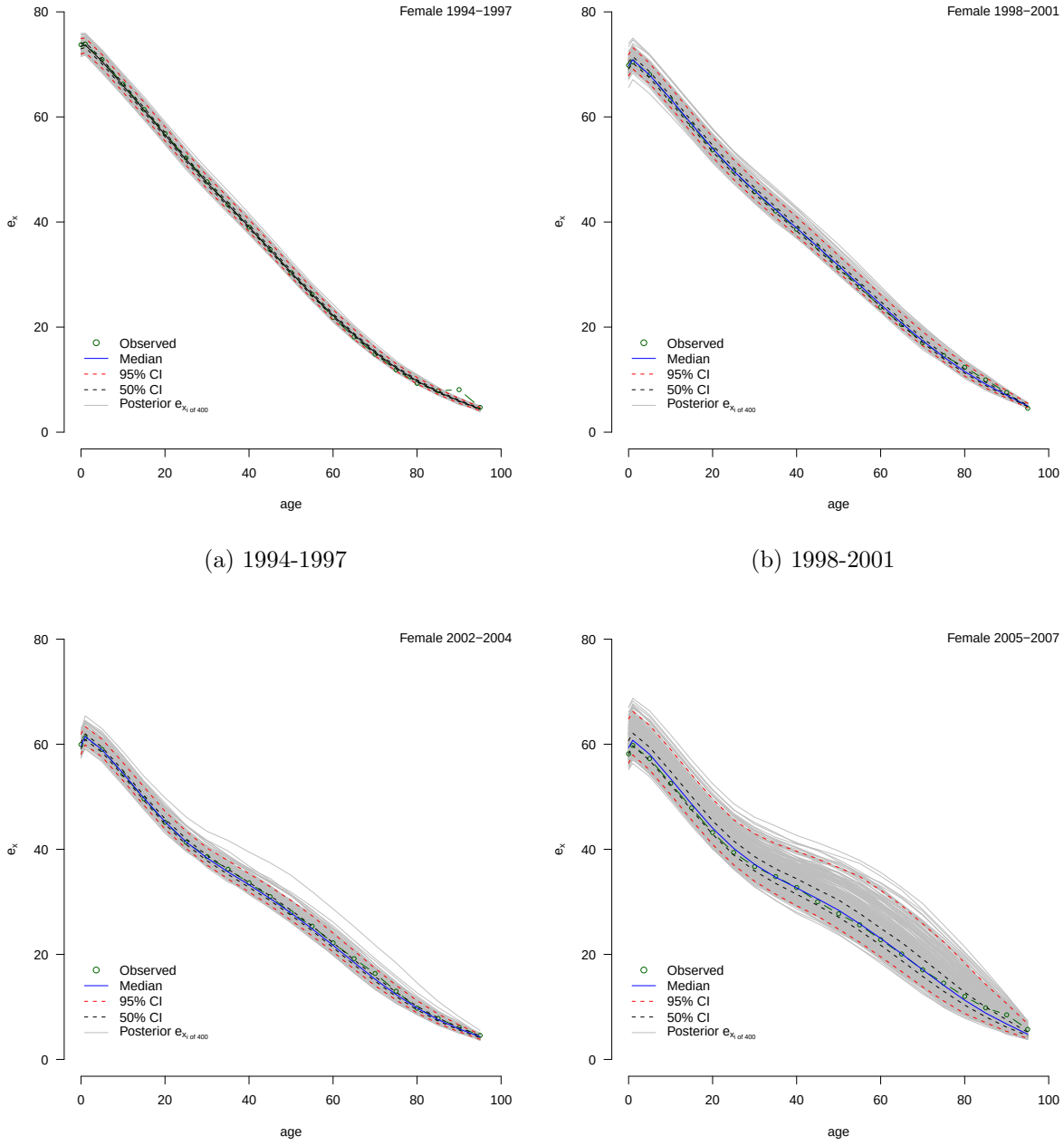
(b) 1998-2001



(c) 2002-2004

(d) 2005-2007

Figure 2.15: Predictive Distributions of Male Age-Specific Expectations of Life for Agincourt HDSS. e_x refers to the expectation of life at age x. Predictive distributions result from fitting the periods with the Heligman-Pollard model estimated with Bayesian Melding.



(a) 1994-1997

(b) 1998-2001

(c) 2002-2004

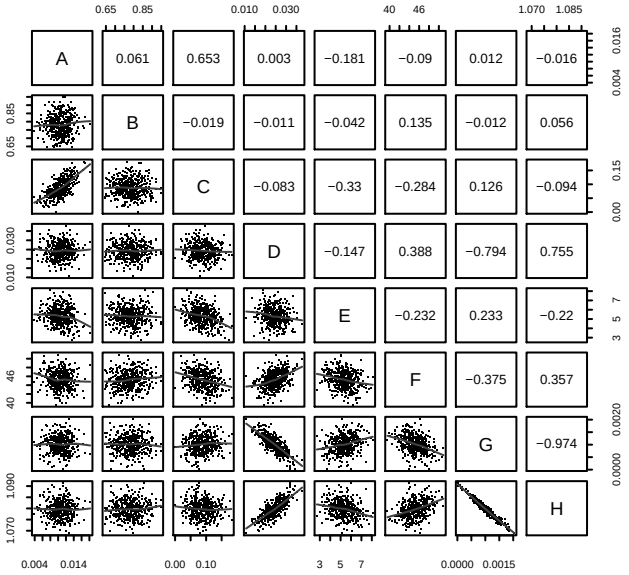
(d) 2005-2007

Figure 2.16: Predictive Distributions of Female Age-Specific Expectations of Life for Agincourt HDSS. e_x refers to the expectation of life at age x. Predictive distributions result from fitting the periods with the Heligman-Pollard model estimated with Bayesian Melding.

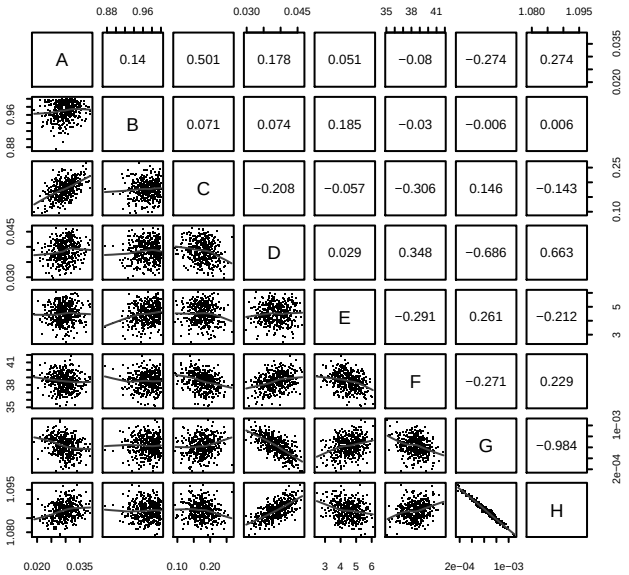
Appendix D: Pairs Plots of Estimated Heligman-Pollard Model Parameters

Correlation in above-diagonal panels, lowess in below-diagonal scatterplots

Males 1994–1997

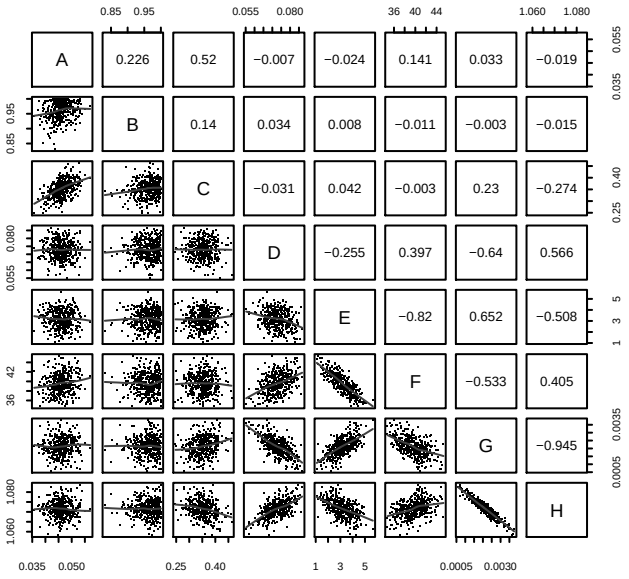


Males 1998–2001

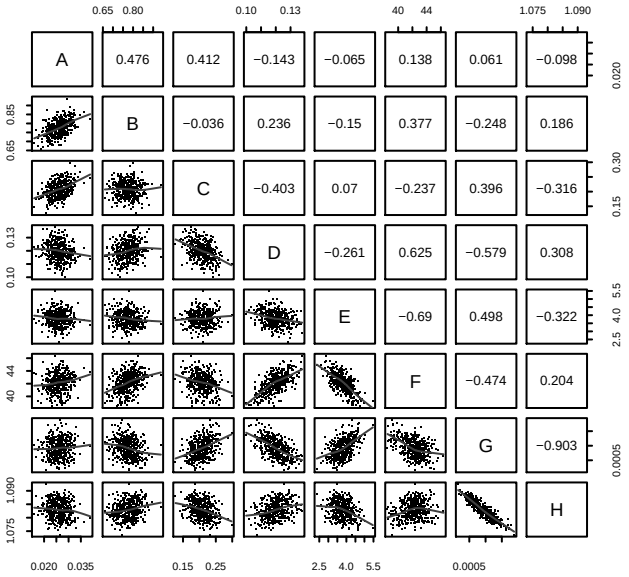


Correlation in above-diagonal panels, lowess in below-diagonal scatterplots

Males 2002–2004

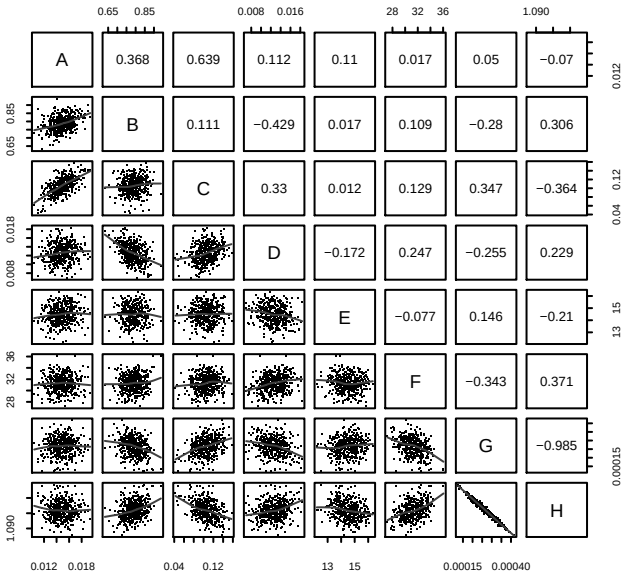


Males 2005–2007

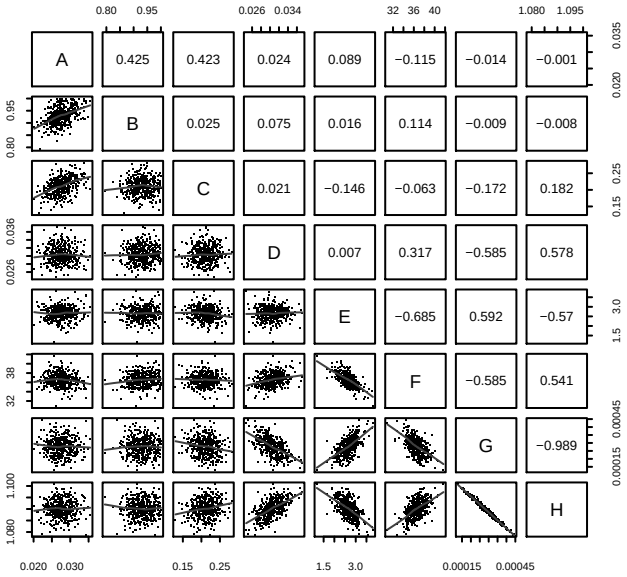


Correlation in above-diagonal panels, lowess in below-diagonal scatterplots

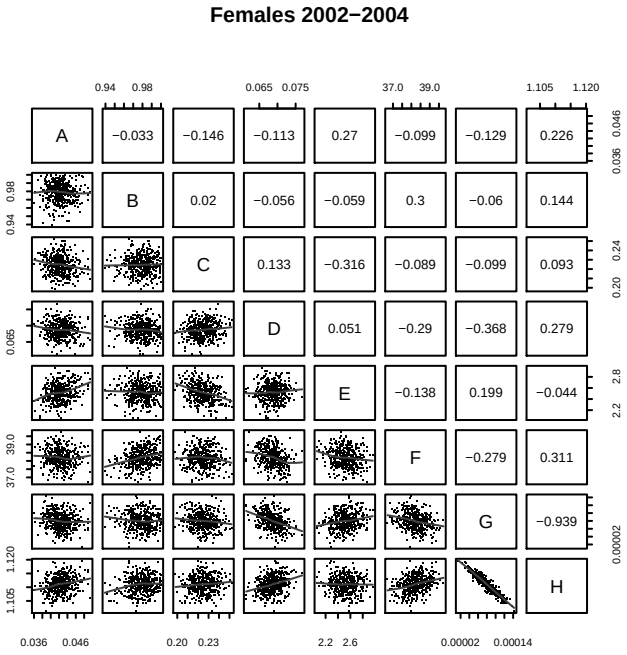
Females 1994–1997

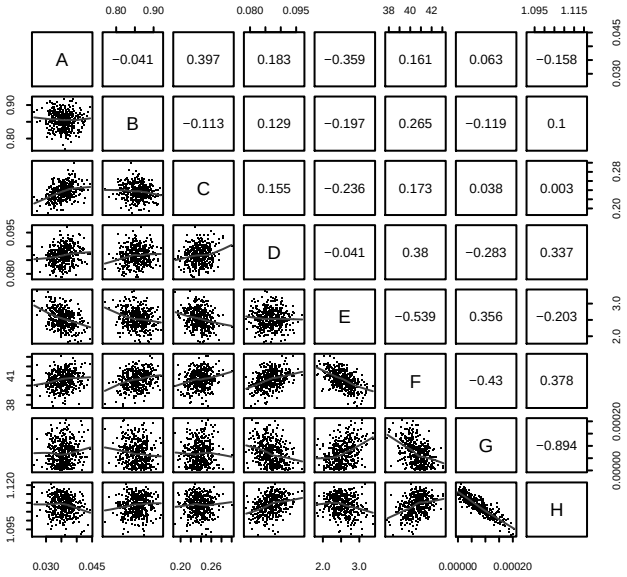


Females 1998–2001



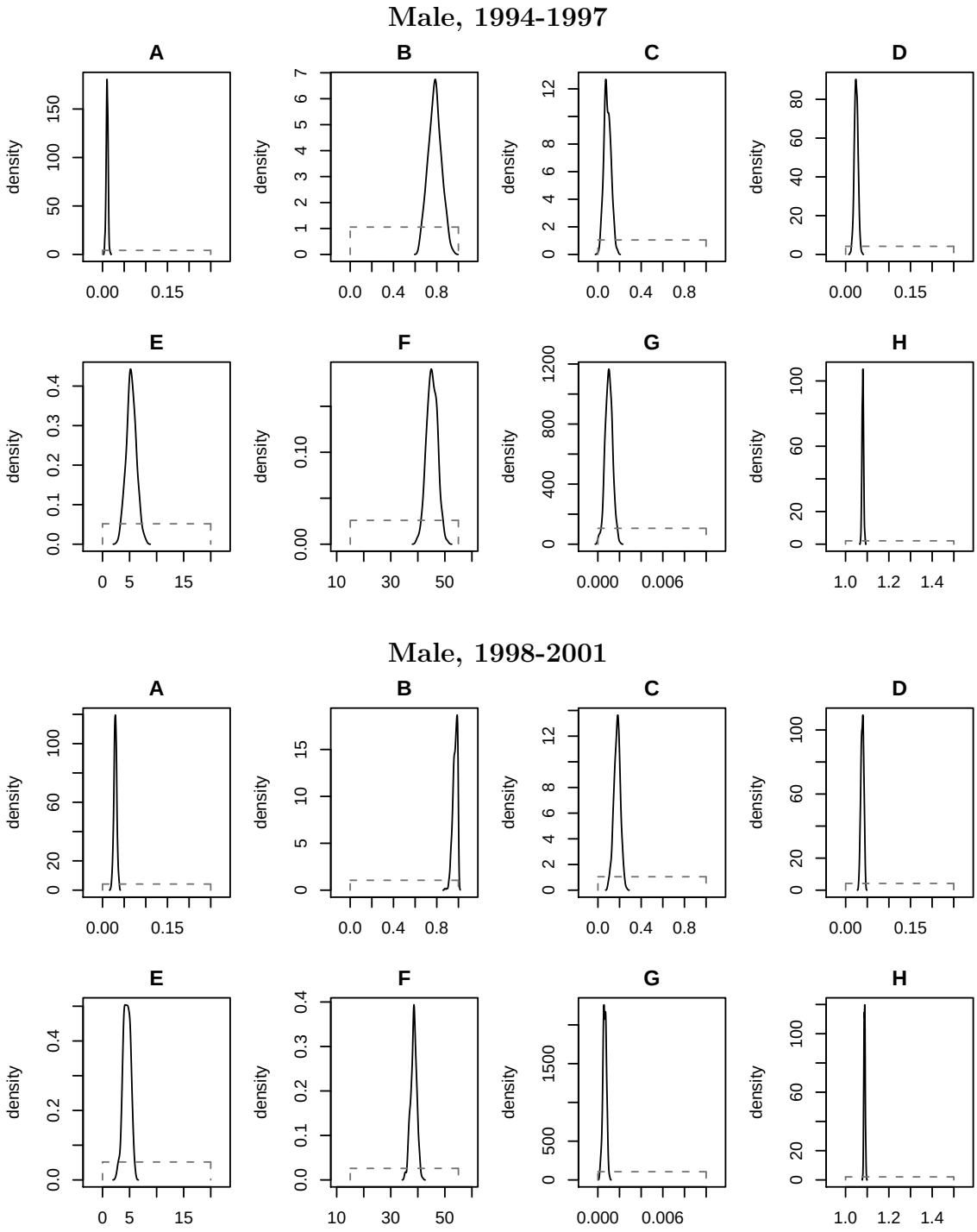
Correlation in above-diagonal panels, lowess in below-diagonal scatterplots



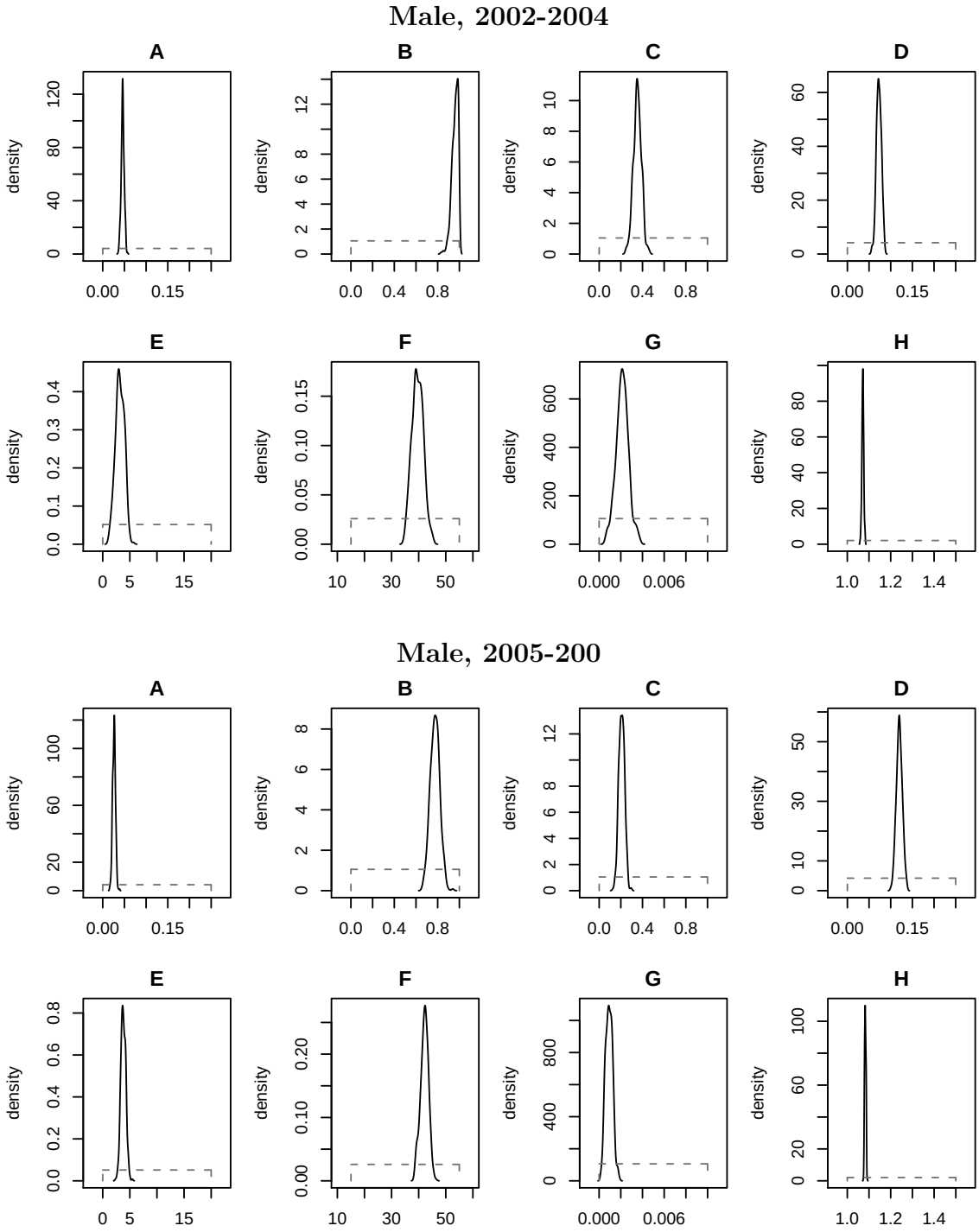


Appendix E: Prior/Posterior Densities for Estimated Heligman-Pollard Model Parameters

Key: Prior Distribution - grey, dashed; Posterior Distribution - black, smooth

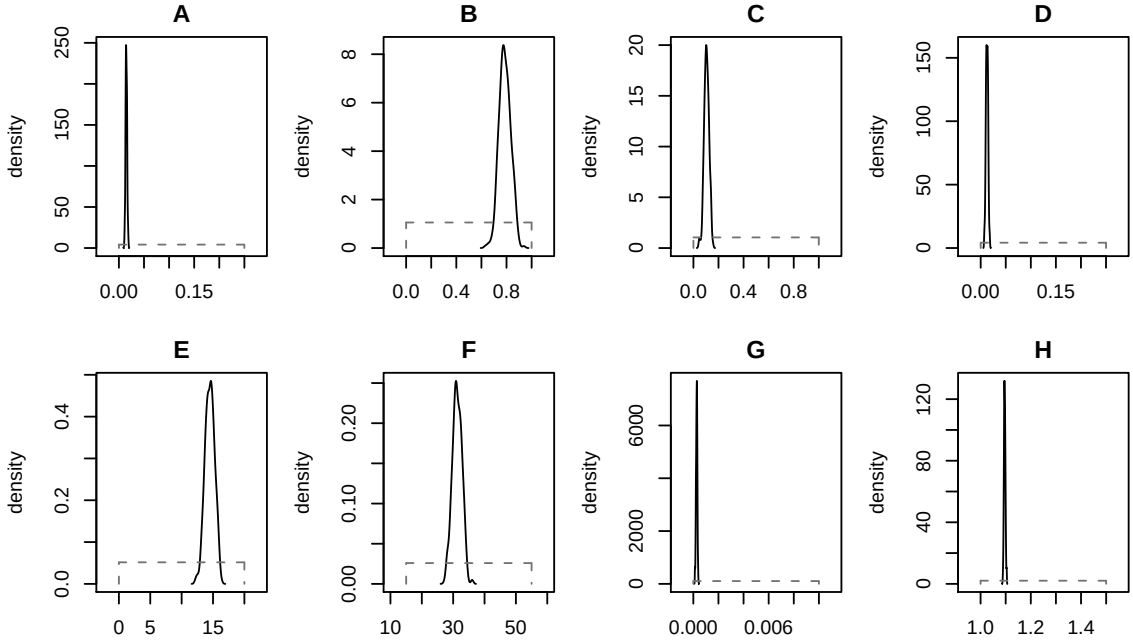


Key: Prior Distribution - grey, dashed; Posterior Distribution - black, smooth

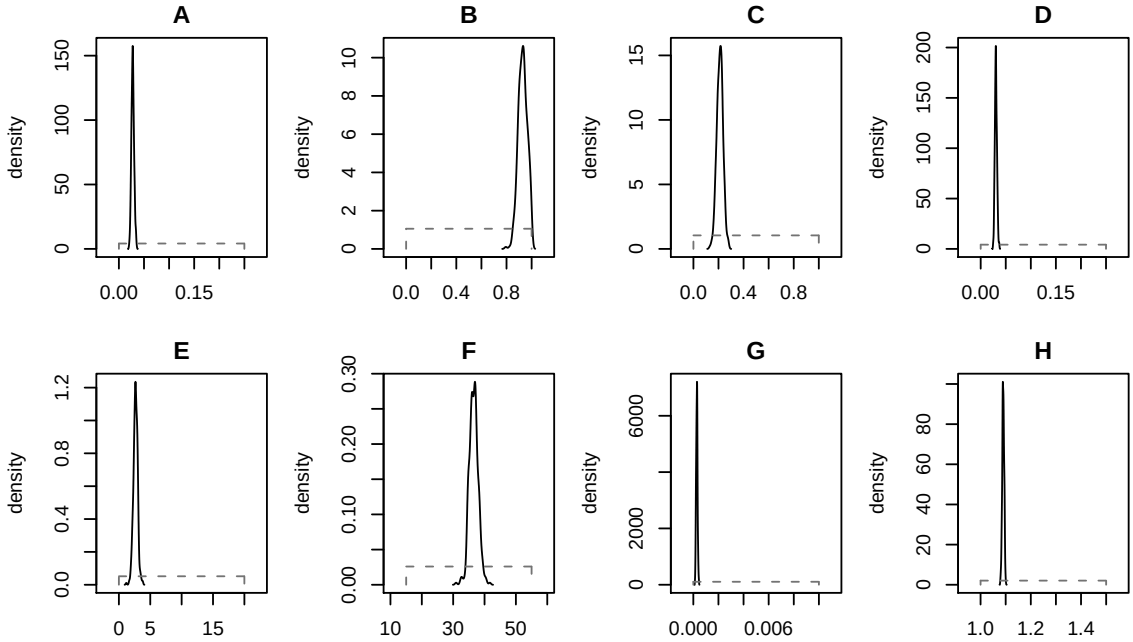


Key: Prior Distribution - grey, dashed; Posterior Distribution - black, smooth

Female, 1994-1997

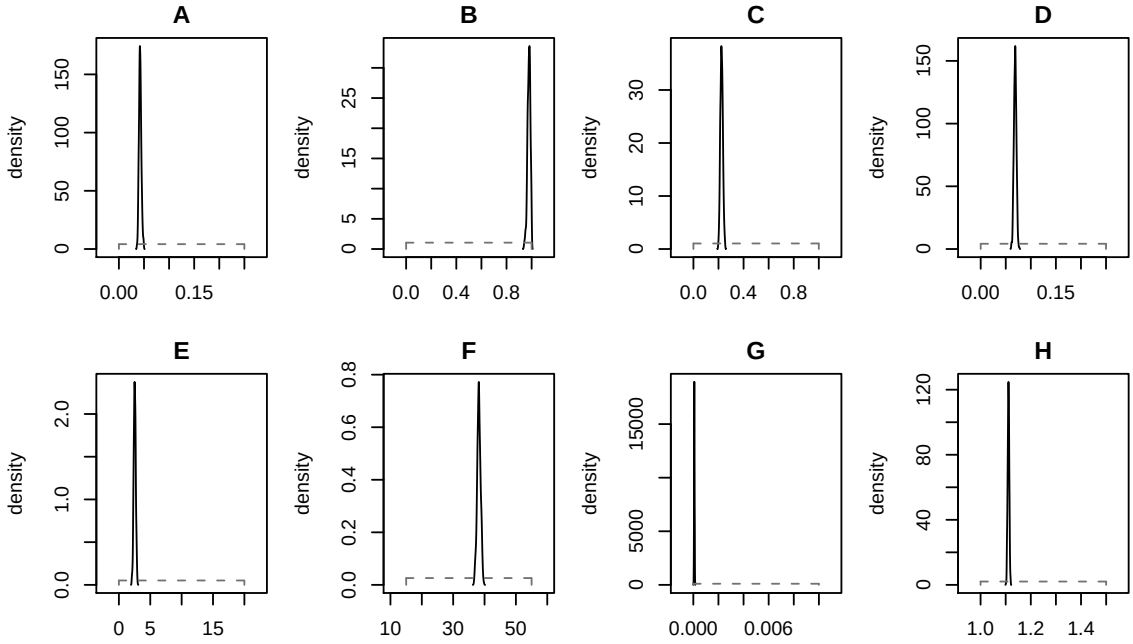


Female, 1998-2001

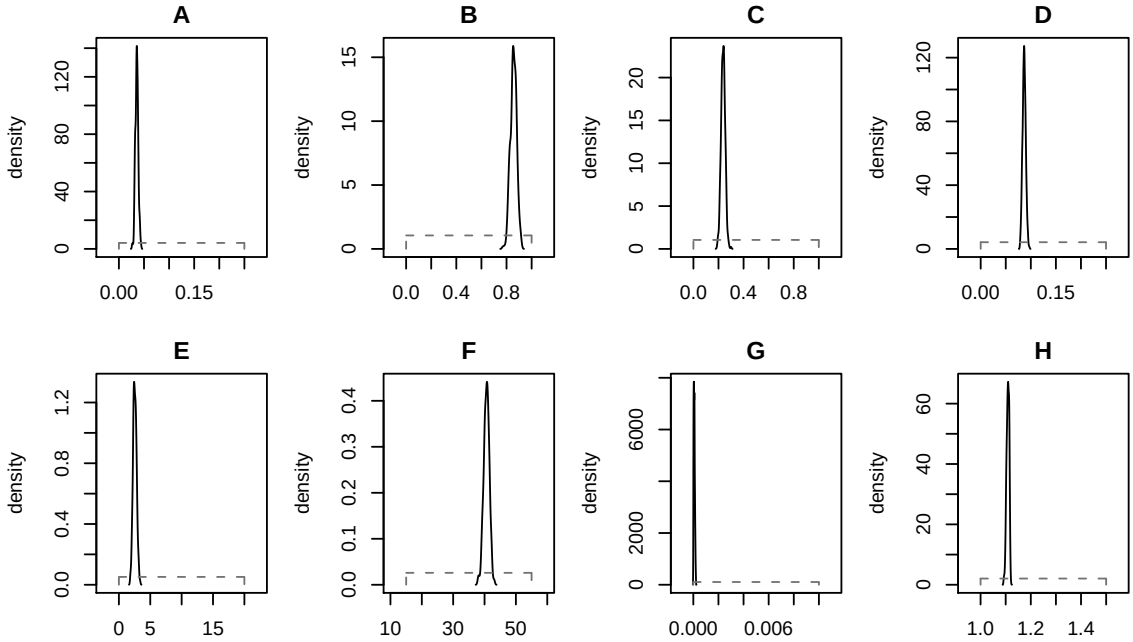


Key: Prior Distribution - grey, dashed; Posterior Distribution - black, smooth

Female, 2002-2004

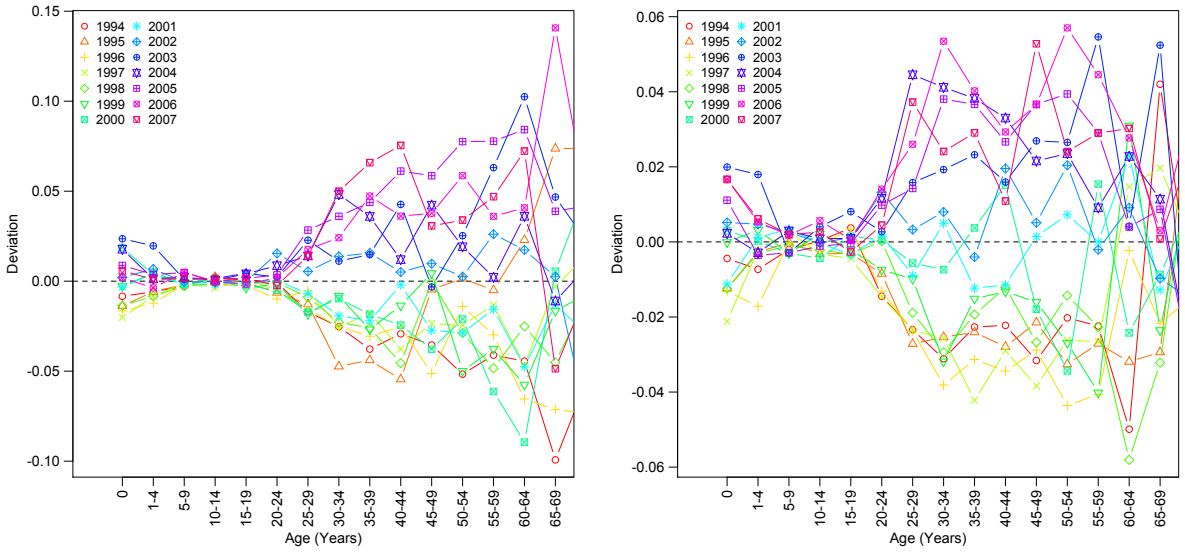


Female, 2005-2007



Appendix F: Period Aggregation

There were two motivations for creating these time intervals. First, I have 14 years of data which I would like to aggregate into intervals that would reduce noise to the greatest extent but also maintain the level and shape of mortality in the individual years and be roughly similar in interval length for the sake of comparison. I first plotted the age-specific deviations between each year's set of age-specific probabilities of death, ${}_nq_x$, and the mean ${}_nq_x$ among all years (figure 2.17). Because the aim is to assess changes resulting from a growing HIV epidemic, I focus on differences in the adult years where HIV is most influential. Panel a of figure 2.17 shows a clear delineation between 2001 and 2002 for males when HIV prevalence increases substantially. While this break is not as clear for females most of the 2002 deviations remain positive while many of the 2001 deviations during the adult years fall below the dashed line. For the sake of having time intervals with similar widths, I next divide the two halves (pre-2002 and 2002+) in half yielding the periods under study. These intervals also contain a similar number of persons at risk of death in the first age-group of the life table, l_0 (see section 3.2).



(a) Male

(b) Female

Figure 2.17: Deviations from mean ${}_nq_x$ schedule for Agincourt HDSS 1994-2007

Appendix G: Likelihood specification for HP model

The data used in this chapter are counts of deaths and persons at risk of dying in each time-sex-age cell thus I use the binomial likelihood which is a natural choice given that the observed ${}_nq_x$ are probabilities. The likelihoods for each age are treated as independent and combined to produce a total likelihood. The likelihood is the probability of witnessing the observed age distribution of deaths given the age-specific probabilities of dying produced by the Heligman-Pollard model with a given set of parameter values. The likelihood is

$$\mathcal{L}\left(\mathbf{y} = (\mathbf{d}, \mathbf{l}) \,|M(\theta) = \mathbf{q}\right) = \prod_{\text{age } i=0}^{100+} \binom{l_i}{d_i} q_i^{d_i} (1 - q_i)^{l_i - d_i} \,. \tag{2.3}$$

where $\mathbf{y} = (\mathbf{d}, \mathbf{l})$ are the data consisting of counts of deaths and persons at risk of dying by age and $M(\theta) = \mathbf{q}$ are the age-specific probabilities of dying output by the Heligman-Pollard model with a given set of values for the parameter vector θ. For a given age group i, l_i is the number of people at risk of dying from the data and d_i is the number of deaths from the data. q_i is the probability of dying in age group i from the model with a specific set of parameter values.

Appendix H: IMIS algorithm

Bayesian melding has been used to analyze other deterministic models (e.g. Alkema et al., 2007; Ševčíková et al., 2007) using the Sampling Importance Resampling (SIR) algorithm (Rubin, 1987; Poole and Raftery, 2000) to sample from the posterior distribution. In this chapter, I instead use incremental mixture importance sampling (IMIS) to obtain samples from the posterior distribution. Incremental mixture importance sampling was originally introduced by Steele et al. (2003, 2006) and was later developed for posterior distributions of continuous parameters, in particular for deterministic demographic-epidemiological models, by Raftery and Bao (2010). The SIR algorithm draws a large number of samples from the prior distribution of the model parameters, weights each sample by its likelihood and then resamples with replacement using the computed weights (Rubin, 1988). Sometimes SIR performs poorly and few distinct values are present in the final sample because a small number of large importance weights dominate the resampling step. The IMIS algorithm addresses this problem. After calculating the likelihoods and weights as in the SIR algorithm, at each iteration a multivariate normal distribution centered at the parameter vector value with the highest importance weight and with a variance-covariance matrix estimated from the weighted covariance of the inputs with the smallest Mahalanobis distances to the center is added to the current importance sampling distribution, thus forming a mixture and successfully representing parts of the parameter space that might otherwise be missed.

If the initial sample from the prior misses a high probability region, IMIS may also fail to cover that region. To address this problem Raftery and Bao (2010) suggest inserting an optimization step after the initial stage (i.e. drawing the samples from the prior distribution and calculating their weights) to yield a mixture of a number of multivariate normal distributions centered around a local maximum of the target distributions. New inputs drawn from the multivariate normal distributions from the optimization step are combined with the initial inputs and new likelihoods and weights are calculated. In the final resampling stage, IMIS resamples x number of inputs with replacement from the mixture distribution with the calculated weights, and it is this final sample that approximates the posterior parameter distribution.

Chapter 3

MODEL LIFE TABLES AND MORTALITY TRANSITIONS FOR DEVELOPED COUNTRIES: AN APPLICATION OF MODEL-BASED CLUSTERING

3.1 Introduction

In the absence of high quality vital registration systems that could accurately describe the age pattern of mortality, many countries must rely on *model life tables* – models that produce (interpolate/extrapolate) complete age-patterns of mortality from age-restricted mortality indicators such as under five mortality (the probability a newborn will die before reaching age five denoted ${}_5q_0$) or adult mortality (the probability a 15 year old will die before reaching age 60 denoted ${}_{45}q_{15}$). In this chapter, I describe a new method for constructing model life tables, which performs better than the best fitting existing model life table systems.

Building on earlier work (Clark, 2002; Clark et al., 2009), I aim to identify commonly observed age-patterns of mortality in the Human Mortality Database (University of California, Berkeley and Max Planck Institute for Demographic Research, Data Downloaded November 2009), which contains data from mostly developed countries, build an easy-to-use model life table system based on those observed mortality patterns, and describe the temporal and regional changes in mortality contained in the Human Mortality Database (HMD). This data source is ideal to develop and validate this model life table construction method and to describe temporal and regional mortality patterns because the data are high quality, well documented, and uniform in age format—requisite features of any database used to generate a model like the one presented here.

Typical age patterns are identified with a model-based clustering method. This clustering method is effectively fully automated allowing for as few subjective decisions as possible on the part of the analyst and without input, permits the patterns to reveal themselves. Using each cluster as the basis for a 'family' in a traditional system of model life tables, a one-parameter model is created to generate 'levels' of mortality within each

family. The result is a new, effectively two-parameter system of model life tables for the countries and time periods included in the HMD.

Although a model life table system for developed countries, where data are typically reliable and available, may seem futile, these models have other uses aside from their primary purpose of extrapolating age-restricted indicators to complete sets of age-specific mortality rates. The analyst can employ model life tables to smooth noisy data or obtain mortality estimates for age intervals that are different from the intervals in available data. Demographers also use model life table systems to estimate historical mortality trends when data is unavailable, a nontrivial matter for analysts interested in population projection. In addition, data from developing regions remains limited (Mathers et al., 2005; Gerland, 2009) and model life table systems generated from developed country data are still essential for mortality estimation in much of the world.

Existing model life table systems were generated using data from restricted geographical locations and time periods, with the commonly used systems dating back some 30 years. In some instances, these systems are unable to reflect contemporary mortality experiences, for example the extremely low child mortality observed in some contemporary developed world settings (Coale and Guo, 1989; Wilmoth et al., 2012) or elevated adult mortality resulting from unusual causes such as HIV. For these reasons, we ought to prioritize updating and improving demographic methods and models including those for use in indirect estimation of mortality. While development of a model life table system using the HMD is consistent with that imperative, ultimately the method developed in this chapter should be applied to what high quality data do exist throughout Africa and other parts of the developing world to advance a model based on data originating in those places where model life tables are still most germane.

In addition to the introduction of a new method, I also draw substantive conclusions about the temporal and regional change in mortality for the countries and time periods in the HMD based on the clusters that emerge from the data. Life tables in the HMD cluster primarily along a temporal dimension with diverse regional representation in each cluster, a distinct finding among model life table systems, which often group similar mortality profiles by geographic region. Each of the resulting clusters of life tables can be thought of as representing a particular cause of death profile and set of epidemiological

conditions that mirrors a stage of the epidemiological or mortality transition as described by Omran and others (Omran, 1971; McKeown, 1976; Olshansky and Ault, 1986; Salomon and Murray, 2002).

I begin with a brief review of the use and structure of existing, commonly used model life table systems, followed by a description of the Human Mortality Database. I then present a detailed description of the model, fitting method, and resulting 'families' followed by a description of the usage and validation of the model. Finally, I discuss future directions for this line of research.

3.1.1 Model Life Table Systems

All model life table systems, including the ones presented in this and the following chapter, are generated from an analysis of a large collection of historical mortality profiles. Unlike mathematical models such as Gompertz (1825) or the all-age model advanced by Heligman and Pollard (1980) that depict the shape of human mortality with relatively few parameters, *tabular* models like those developed by Coale and Demeny (1966) and the United Nations (1982) as well as relational models like Brass (1971) and Murray et al. (2003) contain a larger number of 'true' parameters. However, these systems are able to reflect the detailed variation in human mortality with a very small number of 'effective' parameters as is the case with the model presented in this chapter. A small number of 'effective' parameters is sufficient because most of the parameters become fixed leaving just one (often a measure of child mortality) or two that vary to capture the variation in human mortality.

Because model life table systems are generated from a collection of empirical tables, they can only reflect the extent of mortality variation contained in the empirical data used to generate the system. For instance, the widely used Coale and Demeny and UN model life tables for developing countries cannot generate schedules for which child mortality is extremely low because that kind of mortality profile did not exist at the time those systems were created (Coale and Guo, 1989; Wilmoth et al., 2009). Figure 4.1 plots the relationship between adult and child mortality for the HMD collection (each black circle represents one table) and the relationships produced by the various patterns in those two systems. These figures show that mortality profiles with extremely low child mortality

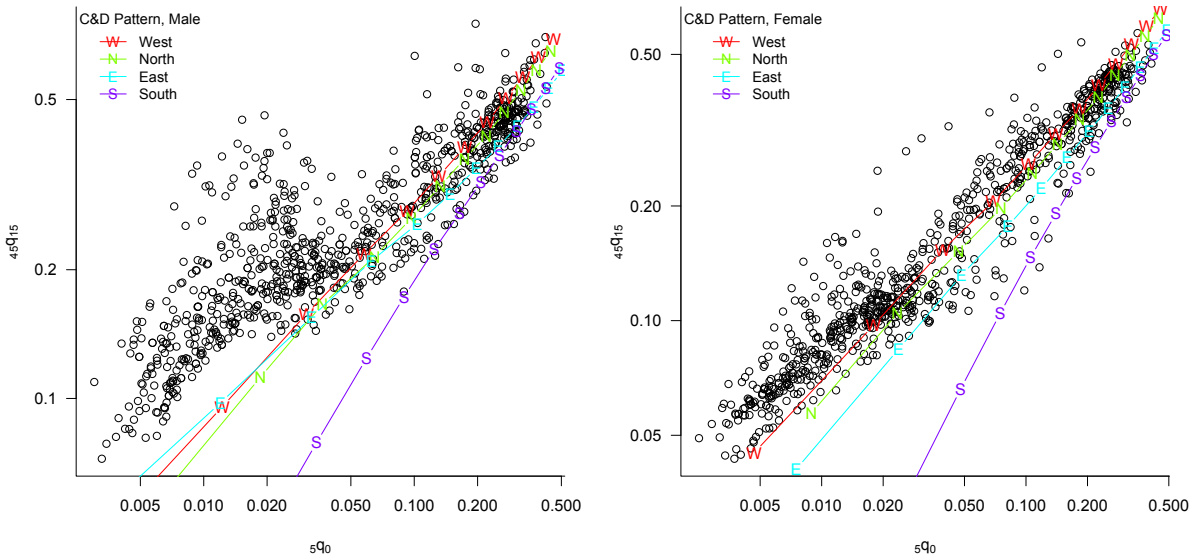
are not well represented by any of the patterns in these two systems. The inability of these models to represent modern mortality schedules is not the result of methodological shortcomings but rather a reflection of the state of mortality at the time. In fact, more recent models calibrated with data from recent time periods such as the model advanced by Wilmoth et al. (2012) are better able to capture this low child mortality situation as shown in figure 4.2 (this figure plots the output from the Wilmoth et al. model after varying the second input parameter 'k'). This issue is poised to become increasingly important as child mortality continues to fall throughout the world (Wilmoth et al., 2012). Furthermore, although not well represented in the dataset analyzed in this chapter, no system to date has included adequate data from Africa or data that could reflect the mortality profile that is characteristic of high HIV prevalence settings. Chapter 4 of this dissertation builds such a system with data from health and demographic surveillance sites throughout Africa and Asia (INDEPTH, 2011).

Because model life table systems are still used extensively to estimate mortality in developing regions as well as for a host of other uses, it is clear that developing a system based on contemporary data that can reflect the widest range of modern mortality experiences is an important improvement over current widely-used systems, but the work presented here has other advantages as well. Typically, the user must supply specific mortality measures like ${}_5q_0$ as input parameters. This method allows the user greater flexibility in terms of the input parameters in that virtually any measure of mortality (${}_1q_0$, ${}_5q_0$, ${}_{45}q_{15}$, e_0, etc.) or available set of age-specific mortality rates either complete or partial can be used to select into the appropriate 'family.'

3.2 Data

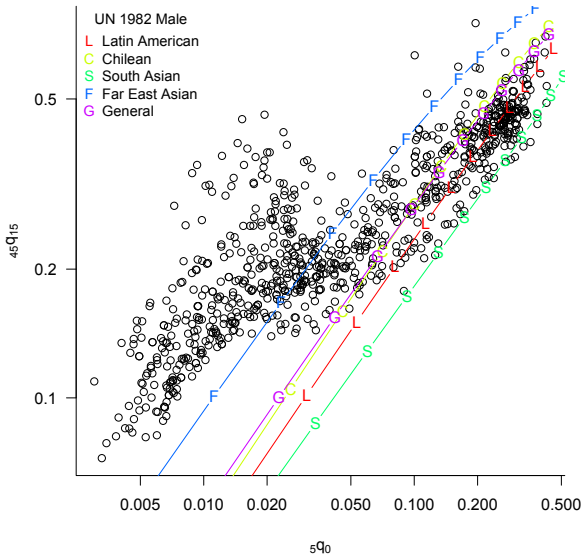
I identify similar age patterns of mortality within a collection of 844 (for each sex) period life tables from the Human Mortality Database. This collection is a publicly available dataset maintained by the University of California, Berkley and Max Planck Institute for Demographic Research (www.mortality.org) and has many advantages for generating a model life table system including the high quality of the data itself and the standardized age and time interval formats.

The HMD contains mortality profiles from 37 mostly developed world regions with

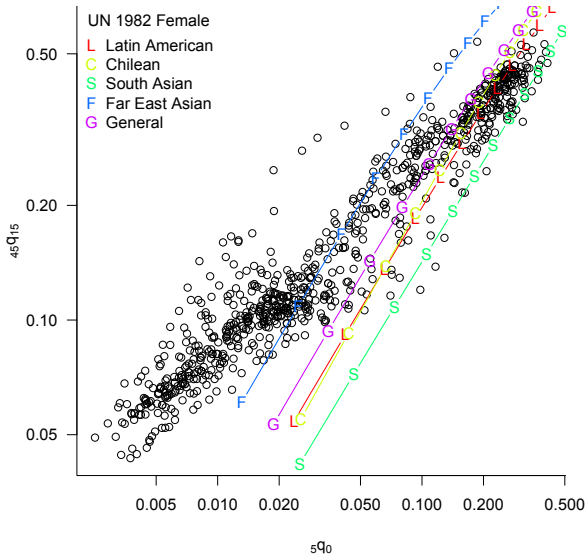


(a) C&D male

(b) C&D female

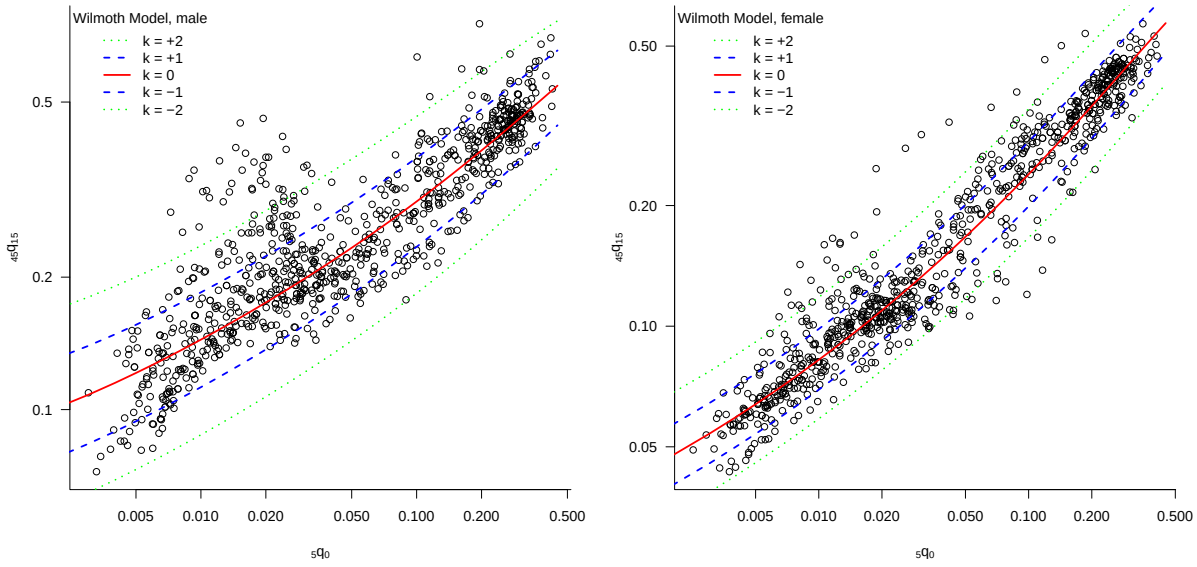


(c) UN male



(d) UN female

Figure 3.1: Relationship between child and adult mortality, observed HMD data (n=844) and Coale-Demeny and UN regional model life table patterns. $_5q_0$ is the probability a newborn will die before reaching age 5 and $_{45}q_{15}$ is the probability a 15 year old will die before reaching age 60.

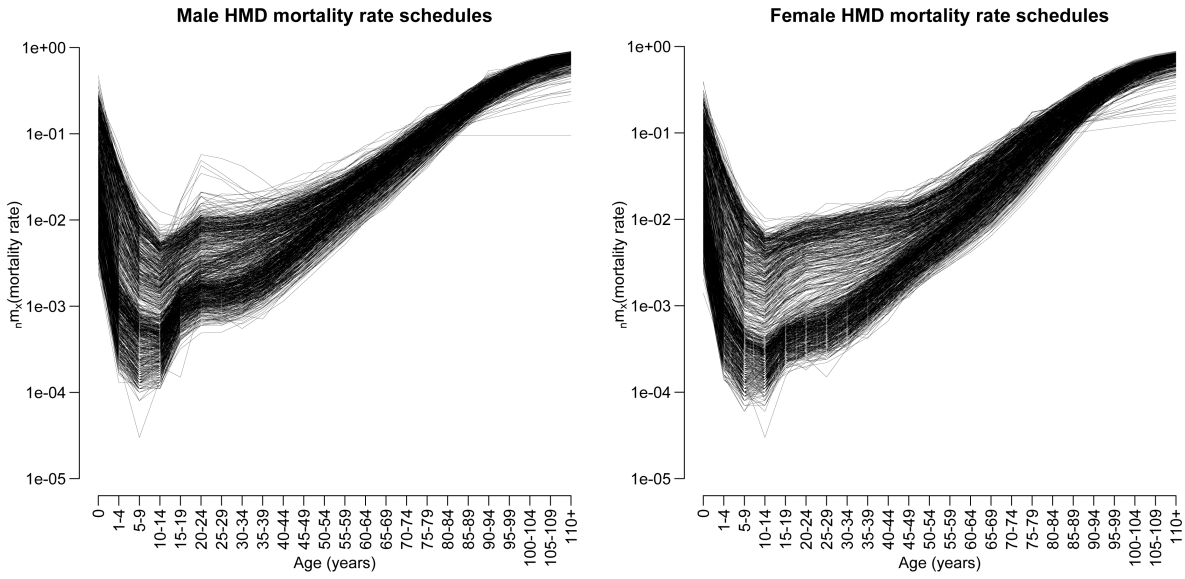


(a) Male (b) Female

Figure 3.2: Relationship between child and adult mortality, observed HMD data (n=844) and output from Wilmoth model resulting from varying the 'k' parameter. ${}_5q_0$ is the probability a newborn will die before reaching age 5 and ${}_{45}q_{15}$ is the probability a 15 year old will die before reaching age 60.

the earliest tables dating back to the mid 18$^{\text{th}}$ century and the most recent from 2007. All of these life tables have been computed from directly-observed deaths and population counts without adjustment except at older ages (Wilmoth et al., 2007) and each table covers a five year period. The life tables in this collection contain approximately 84 million male and female deaths and roughly 5 billion male and 5.5 billion female person years of observation. An important advantage of this dataset is that each table has an open age interval at 110+. This open interval means the system is able to produce complete sets of mortality rates with the same open interval, which exceeds that of other widely used systems.[1] Similar age patterns of mortality are identified in the log mortality rate schedules (denoted as ${}_nm_x$ which is the mortality rate [deaths/person years] from age x to $x+n$), presented in figure 3.3.

[1]In the widely used current systems the open age interval is typically 85+ or 100+.



(a) HMD data male

(b) HMD data female

Figure 3.3: Age-specific Mortality Rate ($_nm_x$) Schedules in the Human Mortality Database, n=844, y-axis shown in log scale

3.3 Method

I begin this section with a discussion of the mortality model that is used to calculate a complete mortality rate schedule. I then describe the components of the model in two broad steps that correspond to the two parameters of the system - identifying 'families' with cluster analysis and defining 'levels' within 'family' with a one-parameter model.

3.3.1 Mortality Model

Motivation

The motivation for this model is two-fold. The model must represent mortality age patterns in a parsimonious way that helps identify regularities among possibly many empirical age patterns, and it must be able to represent a range of model mortality patterns based on the common patterns that emerge from the empirical data.

The general form of the model will be to represent a mortality age pattern as the weighted sum of two or more independent, age-varying components that represent the age-varying nature of the mortality schedule. To this a constant is added at each age

to take into account the non-age-varying level of the mortality schedule. Any remaining differences between the modeled and observed age patterns are captured with a residual term.

The independent, age-varying components necessary for this model can be easily derived from a Singular Value Decomposition (SVD) of the matrix of observed mortality schedules. The resulting left-singular vectors[2] are the independent components we need, and they have the convenient property of encoding the bulk of the variation among the observed mortality schedules in a small number of components (more on this in section 3.3.2).

Model

Assuming k age groups ($k = 24$ for HMD with age groups 0, 1-4, 5-9, 10-14,... 110+), a k x m matrix $\mathbf{M}$ composed of m column vectors of age-specific mortality rate schedules can be expressed as a weighted sum of a number of components whose shapes encode the fundamental age pattern of human mortality and a wide range of variations on that:

$$\mathbf{M} = \mathbf{SB} + \mathbf{C} + \mathbf{R} \tag{3.1}$$

$\mathbf{S}$ is a k x n matrix whose columns are the n 'components' used in the model (derived from the SVD of all of the empirical mortality rate schedules). $\mathbf{B}$ is a n x m matrix whose columns are coefficients that multiply each component schedule contained in $\mathbf{S}$ to yield the age-varying component of each mortality schedule. $\mathbf{C}$ is a k x m matrix whose columns are constants that are added to the result of the weighted sum to modify each mortality schedule in an age-constant way. Finally, $\mathbf{R}$ is a k x m matrix of residuals that account for the remaining difference between the modeled and empirical mortality schedules. See equation 3.2.

---

[2]For readers who are unfamiliar with SVD, the left singular vectors are similar to the components derived from a principal components analysis.

$$\mathbf{M} = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,m} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ m_{k,1} & m_{k,2} & \cdots & m_{k,m} \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{k,1} & s_{k,2} & \cdots & s_{k,n} \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,m} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n,1} & b_{n,2} & \cdots & b_{n,m} \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} c_{\,\cdot\,,1} & c_{\,\cdot\,,2} & \cdots & c_{\,\cdot\,,m} \\ \vdots & \vdots & \vdots & \vdots \\ c_{\,\cdot\,,1} & c_{\,\cdot\,,2} & \cdots & c_{\,\cdot\,,m} \end{bmatrix} \tag{3.2}$$

$$\mathbf{R} = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k,1} & r_{k,2} & \cdots & r_{k,m} \end{bmatrix}$$

Ignoring the residuals, $\mathbf{SB} + \mathbf{C}$ represents the modeled mortality schedules. $\mathbf{SB}$ captures the age-varying component of the mortality schedules and $\mathbf{C}$ represents the non age-varying level of each mortality schedule.

Equation 3.3 presents the model when just a single mortality schedule is being calculated. The component vectors $s_{(\cdot,\cdot)}$ can be thought of as a new basis in age-space with the special property that most of the variation in the data is represented by a small subset of the these. The weights $b_{(\cdot)}$ determine a given point in this space, and then c adds a constant amount to each age.

If only one mortality schedule is involved:

$$\begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{bmatrix} = b_1 \cdot \begin{bmatrix} s_{1,1} \\ s_{2,1} \\ \vdots \\ s_{k,1} \end{bmatrix} + b_2 \cdot \begin{bmatrix} s_{1,2} \\ s_{2,2} \\ \vdots \\ s_{k,2} \end{bmatrix} + \quad \cdots \quad + b_n \cdot \begin{bmatrix} s_{1,n} \\ s_{2,n} \\ \vdots \\ s_{k,n} \end{bmatrix} + \begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix} + \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{bmatrix} \tag{3.3}$$

The effective parameters in this model are the b's, and as mentioned above, when only a small number of age varying components is necessary to account for the bulk of the variation in the empirical data set, then the number of b's necessary is small, making the model effectively parsimonious. Of course in reality the model is much less parsimonious because the component vectors $\mathbf{S}$ are parameters themselves, albeit fixed. Provided the

original data set from which they are calculated is highly varied, they will be capable of representing a wide range of age variation and can be thought of as permanently fixed.

3.3.2 Identifying Similar Age Patterns of Mortality

A model-based clustering method (Banfield and Raftery, 1993; Fraley and Raftery, 2002, 2006) is used to identify similar age patterns of mortality. Before the cluster analysis, I first reduce the dimensionality of the data using Singular Value Decomposition (SVD). In order to maintain congruence of periods and locations between male and female schedules, I bind the male and female schedules and perform the SVD on the resulting 48×844 matrix. SVD decomposes a matrix into three smaller matrixes including one whose columns are orthogonal and point in the directions with most variation in the original space – the *left singular vectors* (LSV). These vectors concentrate the information in the original matrix into a smaller number of dimensions, and as a result the information in the original matrix can be represented by just the first few left singular vectors. Thus, the dimensionality of the data can be reduced from 24 (or 48) age groups to just four or five component vectors. Figure 3.4 plots the first five component vectors (the LSV) from the SVD of the HMD mortality age profiles. One can see from this figure that the first component captures the underlying shape of mortality with each successive component capturing smaller and smaller variations on that.

The reduction in dimensionality is completed by regressing (simple OLS linear regression) each empirical mortality schedule on the first ten component vectors $\mathbf{S}$ and storing the resulting coefficients and constants in a new data set. These coefficients and constants now represent the bulk of the variation in the original mortality schedules. Keeping ten coefficients provides an opportunity to vary the number of components used in the clustering and choose the number that provides the best clustering.

With this new reduced-dimension data set, I use the model-based clustering method to identify robust clusters. This technique is a nearly fully automated, robust clustering method that identifies the number and shape of clusters that maximizes the Bayesian Information Criteria (BIC). Model-based clustering puts cluster analysis on solid statistical footing and quantifies uncertainty about the results. This method yields both the BIC values for different numbers of clusters and the classification of the data using the number

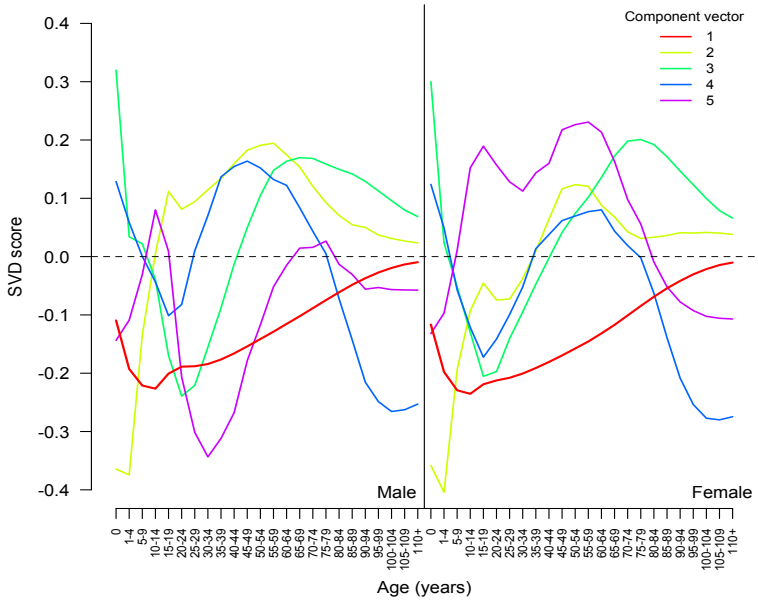


Figure 3.4: First five left singular vectors from the Singular Value Decomposition of the HMD mortality rate schedules. **S** from equations 3.1 and 3.2

of clusters with the greatest BIC value.

A priori there is no objective way to choose how many of the new 'reduced' dimensions (represented by the coefficients) to include in the clustering, so clustering is performed on the reduced dimension data set using 2-10 coefficients, and the clustering classification from each is saved in a new data set. The best clustering is chosen by calculating a new fit metric, the "Total Deviation from Median" (TDM). The TDM is the sum of the absolute differences between each mortality schedule and the median of all mortality schedules in the cluster to which it is assigned. Lower values of the TDM indicate less variation and tighter clustering among mortality schedules in each cluster, and consequently, the best clustering has the lowest overall TDM value. The HMD calibration of this model produces the best clustering with just the first four coefficients/dimensions.

3.3.3 Model Life Tables

Following the structure of existing systems, the one presented here is composed of 'families' with different 'levels' of mortality in each. Each 'family' is based on one of the clusters identified using the procedure described above. The structure of the system is presented

in equation 3.4.

$$\text{underlying family-specific age pattern} + \alpha * \text{family-age-specific deviation} \tag{3.4}$$

where α varies to generate levels within a family.

Life Table Families

The underlying-family-specific age pattern of mortality for each family is identified by first obtaining the median set of coefficients for each family (cluster) and inserting them into equation 3.1 to obtain a 'family' pattern. The resulting cluster-specific median mortality rate schedules are the underlying mortality age profiles on which the model life table families are based $\mathbf{M}_f$, where f indexes the families.

Age-Varying Mortality Levels within a Family

There is variation in the overall level of mortality within each of the clusters that underly the model life table families. This source of variation is included in equation 3.4 with the family-age-specific deviation. Each age-specific deviation is a weighted average of variation from two sources. The first captures age-specific variation within the family as the difference between the 97.5^{th} quantile and the median for age group i and family f when α is positive, and the difference between the 2.5^{th} quantile and the median for age group i and family f when α is negative. The second source is the same as the first but represents age-specific variation within the *entire* HMD dataset; the calculation is the same except all of the HMD life tables are included.

As α approaches an absolute value of 1 and the resulting age pattern moves farther from the family-specific underlying age pattern, more weight is given to the differences calculated from the the entire dataset. When α is 1 the balance is about half-half, and when α moves beyond 1 the 'all-tables' age-specific deviation progressively dominates. Consequently, as the family-specific level approaches the 'edge' of the data for each family, the cluster invariant ('all-tables') age-specific deviation takes over. This method produces smooth family-age-specific deviations for mortality levels approaching, at, and beyond the levels represented by the data within each family, and as a result the model is able to extrapolate reasonably to very low and high mortality levels within all of the families.

The age-specific deviations are represented by the model in the following way. Within a group of life tables (either a family or the whole HMD data set), the age-specific change $\mathbf{D}$ below or above the median is represented by:

$$\text{below the median}: \quad \mathbf{D}_{f_-} = \mathbf{M}_f\,[\text{median}] - \mathbf{M}_f\,[2.5^{\text{th}}\text{ quantile}] \tag{3.5}$$

$$\text{above the median}: \quad \mathbf{D}_{f_+} = \mathbf{M}_f\,[97.5^{\text{th}}\text{ quantile}] - \mathbf{M}_f\,[\text{median}] \tag{3.6}$$

The final expression for a life table in family f at level α is:

$$\text{when } \alpha < 0: \quad \mathbf{M}_f(\alpha) = \mathbf{M}_f + \alpha \cdot \left(\mathrm{e}^{-0.75\cdot|\alpha|} \cdot \mathbf{D}_{f_-} + (1 - \mathrm{e}^{-0.75\cdot|\alpha|}) \cdot \mathbf{D}_{h_-}\right) \tag{3.7}$$

$$\text{when } \alpha = 0: \quad \mathbf{M}_f(\alpha) = \mathbf{M}_f \tag{3.8}$$

$$\text{when } \alpha > 0: \quad \mathbf{M}_f(\alpha) = \mathbf{M}_f + \alpha \cdot \left(\mathrm{e}^{-0.75\cdot|\alpha|} \cdot \mathbf{D}_{f_+} + (1 - \mathrm{e}^{-0.75\cdot|\alpha|}) \cdot \mathbf{D}_{h_+}\right) \tag{3.9}$$

where f indexes families and h is the whole HMD dataset. The minus sign in (f or h)$_-$ indicates age-specific deviations below the median, and the plus sign indicates age-specific deviations above the median. The value of -0.75 for the coefficient in the exponential weight is chosen so that when $|\alpha| = 1.0$ there is approximately equal weight given to the family-specific and all-tables age-specific deviations.

Similar to previous systems, this one indexes the levels within each family on the expectation of life at birth. I use the `optimize()` function in the `R` statistical software (R Development Core Team, 2011) to find values of α such that the expectation of life at birth ranges from 30 to 90 within each family.

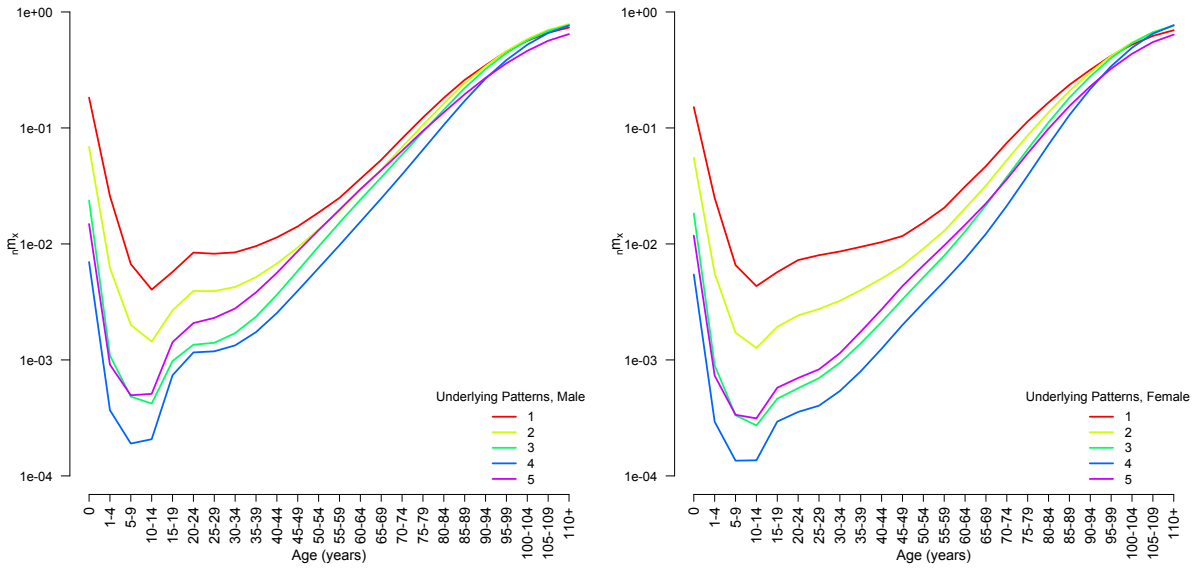
Users of the system can supply any of a variety of measures of child or adult mortality and discriminant analysis can be used to identify the family in which the supplied measure is most likely to fit. Discriminant analysis (DA) is used to classify cases into the values of a categorical dependent variable. DA uses observations of a known classification to classify other observations (Fraley and Raftery, 2006). Once a model is specified, DA can be used to 'predict' which classification or group a particular case most likely belongs.[3] Once the appropriate family has been determined, the user can select a level of mortality or desired life expectancy within the identified family.

[3]The `R` package `Mclust`, which was used to perform the cluster analysis, contains functions to perform discriminant analysis where any portion of the age range or any summary measure like e_0 can be used with the known classification to train the DA model. These functions also provide a likelihood for classification into each cluster, which makes it possible to estimate a full set of mortality rates based on two or more clusters by weighting the underlying family patterns and deviations according to these likelihoods.

3.4 Results

3.4.1 Mortality Patterns

The cluster step identified five distinct mortality patterns for males and females. The underlying family patterns, $\mathbf{M}_f$, are presented in figure 3.5. The left panel of that figure shows that nearly all of the male patterns contain the expected young-adult mortality hump. Patterns 3 and 5 show similar levels of child mortality while diverging somewhat in adult mortality. The rest of the patterns show wide variation in both adult and child mortality for both males and females.

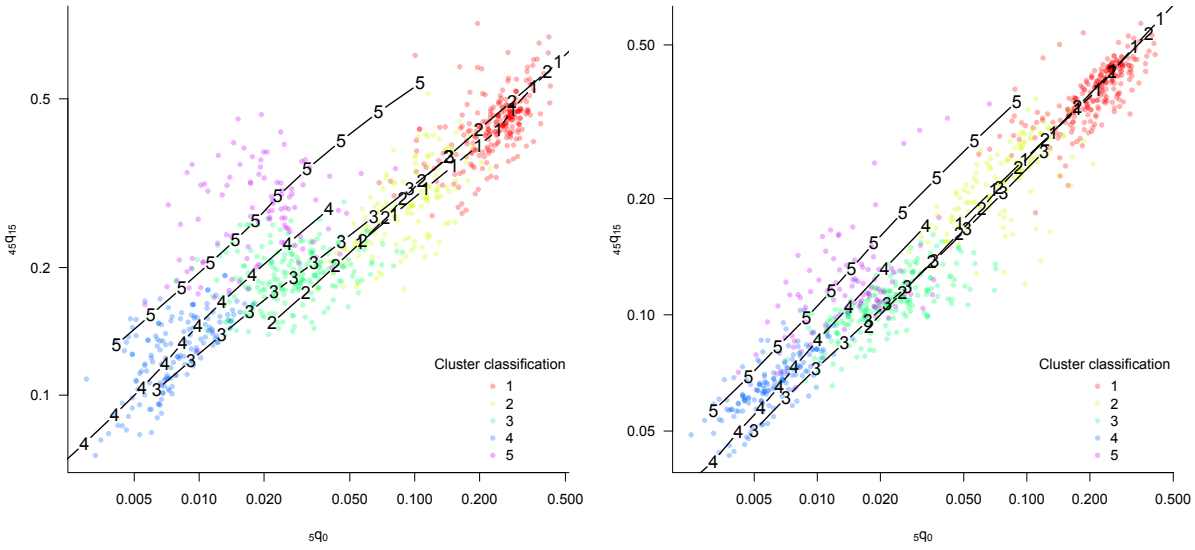


(a) Male

(b) Female

Figure 3.5: Underlying Family Age-Patterns of Mortality, $\mathbf{M}_f$, identified in the Human Mortality Database. y-axis shown in log scale

Figure 3.6 plots the relationship between adult mortality ($_{45}q_{15}$) and child mortality ($_5q_0$) in the observed data (small circles) as well as the resulting relationships from each model family after varying α. This figure reveals that by varying α this system can cover a wide range of human mortality experiences. For example, patterns 4 and 5 cover cases of low child mortality at both high and low adult mortality respectively, while patterns 1 and 2 result from very high adult and child mortality conditions.



(a) Male (b) Female

Figure 3.6: Child versus Adult Mortality in the Human Mortality Database and resulting relationships after varying α parameter (level). α varied from -1 to 1 with black numbers corresponding to family patterns at various levels. Observed data shown as small circles color coded according to cluster classification (family). Each circle represents one table from the HMD.

3.4.2 Components of each Cluster/Family

Each cluster or family contains life tables from a wide range of geographic locations. In contrast, the life tables in each cluster tend to represent similar historical periods. Figure 3.7 shows the distribution of cluster membership over time for each country/region. Each circle represents one life table from the HMD, which is color coded according to the classification from the clustering step. The countries in this figure are ordered along the vertical axis according to their United Nations Statistics Division Standard Country and Regional Codes classification except for former Soviet states, which exhibit a unique pattern (shown at the top of the plot). This figure makes clear the temporal pattern in the clustering results and demonstrates that nearly every country/region transitions from a high mortality experience (pattern 1 in red) to the lowest and most recent mortality profile to emerge, pattern 4 (blue circles).

The transition between patterns through time mirrors the stages of epidemiological

and health transition described by Omran and others (Omran, 1971; McKeown, 1976; Olshansky and Ault, 1986). Moreover, the decline in the level of mortality over time likely reflects shifts in the composition of the cause of death strucutre. The families identified by the clustering step and the underlying patterns that emerge from them represent these shifts in cause composition. I now turn to a discussion of the cluster membership for each family from earliest to most recent in the mortality transition touching on the differences between sequential patterns in both form and cause of death.

The "Early Transition" Pattern

As one can see from figures 3.5 and 3.6, pattern 1, called the "Early Transition" pattern, is characterized by relatively high mortality at all ages. The elevated mortality exhibited in this pattern most closely resembles the earliest stage of transition where the cause of death structure is dominated by infectious and parasitic diseases resulting in male life expectancy of 43.5 years and female life expectancy of 46 years (when $\alpha = 0$) for this pattern.

Unlike other model life table systems which identify regional patterns among a collection of countries/regions, the patterns emerging from this work may be present in geographically dissimilar areas during any given period. For instance, the "Early Transition" pattern is found consistently throughout northern Europe and, when data are available, during early periods in western and southern Europe. The "Early Transition" pattern can also be found in the oldest data from New Zealand. This is the earliest pattern to emerge (the median starting year for the life tables in this family is 1880) and no other pattern appears until close to the turn of the 19th century.

The "Early-Mid Transition" Pattern

Pattern 2 with a median starting year for all life tables contained in this cluster of 1935 is the next pattern to emerge and is referred to as the "Early-Mid Transition" pattern. It is illustrative to discuss the differences between patterns as countries transition from one to the next through time. Figure 3.8 plots the differences between the age-specific mortality rates for pairs of underlying family patterns. The black line in this figure shows the difference between the "Early Transition" pattern and the "Early-Mid

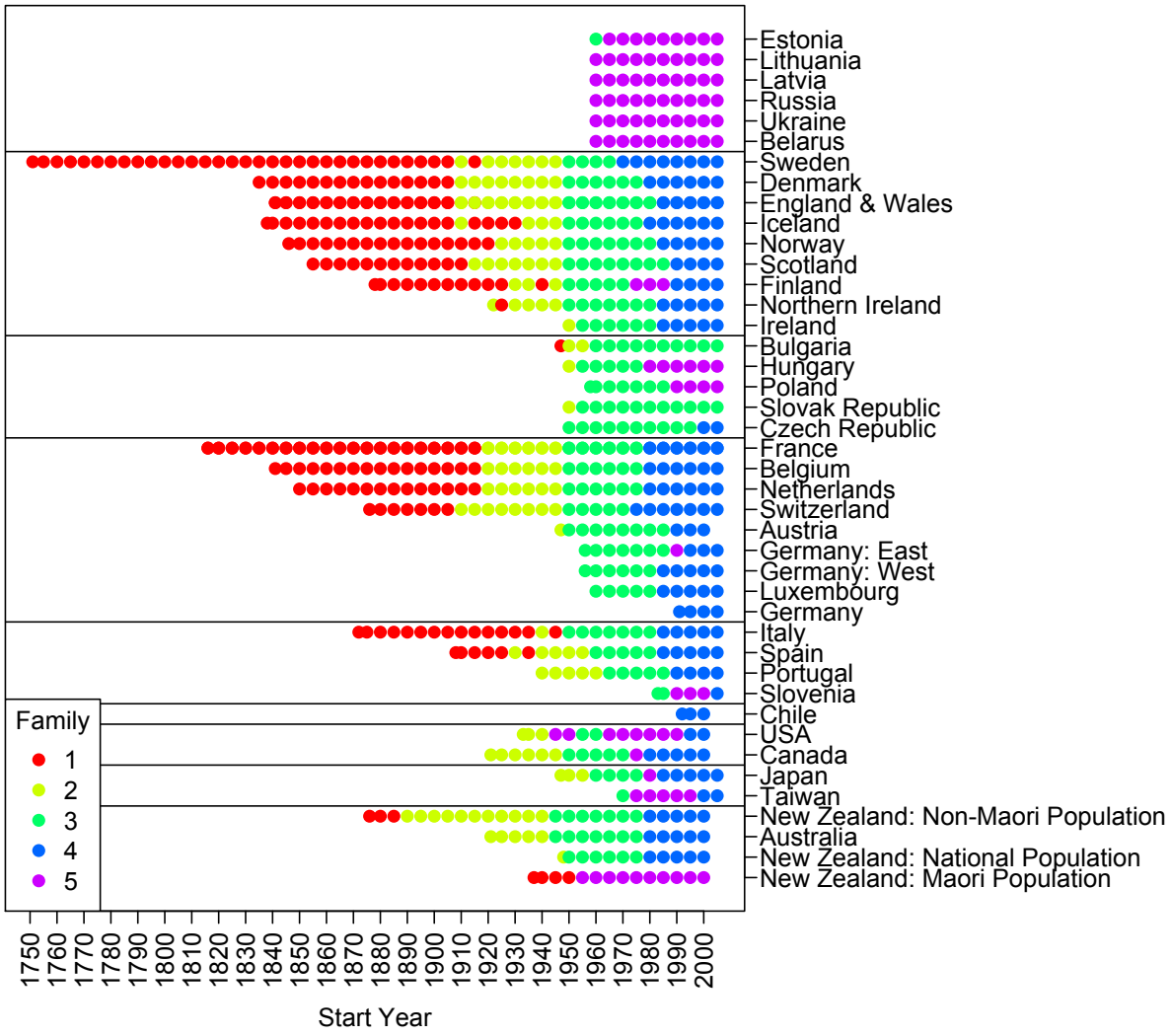
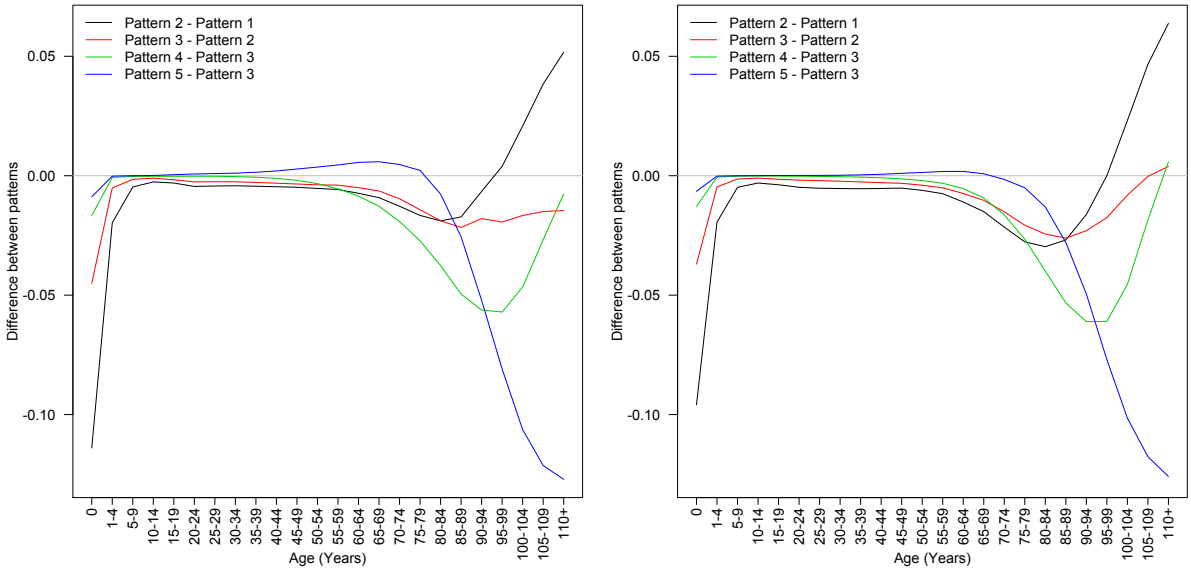


Figure 3.7: Country/Region by Time Period and Family Membership (cluster classification) for HMD data. Countries/Regions are ordered based on UN regional code (except for former Soviet states shown at the top of the figure) and then within UN region by number of family patterns per country. The horizontal lines delineate regions with the regions in the following order from top to bottom: Former Soviet states, Northern Europe, Eastern Europe, Western Europe, Southern Europe, North America, Asia, Oceania.



(a) Male

(b) Female

Figure 3.8: Pairwise differences between underlying family patterns. Pairs of patterns used to calculate differences were selected in chronological order according to the mean starting year for the life tables in each cluster. Pattern 5 is removed from the chronological ordering and the difference shown here is between patterns 3 and 5 as pattern 5 is a slight modification of pattern 3.

Transition" pattern. Although the largest changes in mortality between these patterns come from a steep decline in childhood and infant mortality, the largest improvements in adult mortality between any two patterns are between these two. Notably, mortality above age 95 is somewhat higher for both males and females in the "Early-Mid Transition" pattern than in the "Early Transition." The improvements in child and adult mortality likely originate with declines in infectious and parasitic diseases like influenza, pneumonia, and diarrheal diseases (Omran, 1971; Preston, 1976; McKeown, 1976, 1979). In terms of the life expectancy for the underlying family patterns, there is a 16- and 18-year increase for males and females respectively between these two patterns.

Like the "Early Transition" family, the "Early-Mid Transition" family, contains life tables from countries throughout western, northern, and southern Europe as well as from New Zealand. In addition, this family also contains tables from Japan as well as the oldest tables from the United States and Canada. The range of life table starting years in this family extends from 1890 to 1960 although this pattern does not arise in any significant

way until closer to 1910. Prior to that time it was restricted only to the Non-Maori population in New Zealand. This pattern largely disappears by 1950 save for Spain, Portugal and Japan, which experienced this pattern up until the 1960s.

The "Mid Transition" Pattern

The first instance of pattern 3, the "Mid Transition" pattern, appears in New Zealand and Australia in 1945 but does not reach Europe until 1950. Most of western, northern and southern Europe as well the US, Canada, New Zealand, and Australia transition to a lower mortality profile by about 1975. Notably, two countries, the Slovak Republic and Bulgaria, have yet to transition to the lowest mortality profile and we see the "Mid Transition" pattern up until the latest starting date for this data, 2005. Although now characterized by relatively low mortality, the Czech Republic also experienced the "Mid Transition" pattern until a relatively late 2000.

The red line in figure 3.8 shows the difference in age-specifc mortality rates between the "Early-Mid Transition" pattern and the "Mid Transition Pattern." Similar to the transition from the "Early" to the "Early-Mid" pattern, the "Mid Transition" pattern sees improvements in childhood mortality, although not as steep a decline as before. Contrasting with the increase in old-age mortality from the "Early" to the "Early-Mid" pattern, we see improvements in virtually all age groups for old-age mortality as countries transition to the "Mid" pattern.

The "Post Transition" Pattern

Pattern 4, the "Post Transition" pattern, contains the lowest mortality rates of all patterns that arise from the HMD and appears most recently with a median start year of 1995. This pattern is unlikely to be found in previous systems as this pattern probably did not exist at the time those systems were created. The "Post Transition" pattern can be seen throughout Europe beginning as early as 1970 in Sweden, although this patten is not wide spread until the 1980s. While New Zealand and Australia settle into this pattern beginning in 1980s as do Canada and Japan, for a few, diverse countries including the US, Taiwan and the Czech Republic, this pattern does not emerge until the 1990s.

The "Post Transition" pattern results in the highest life expectancies for the underly-

ing family pattern (when $\alpha = 0$) of roughly 74 for males and 80 for females. The lower overall level of mortality reflects the complete shift from infectious and parasitic disease to degenerative and non-communicable diseases largely impacting mortality at the oldest ages (Omran, 1971; Murray and Lopez, 1996, 1997; Lopez et al., 2006).

The "Modified Mid Transition" Pattern

While most populations have made significant gains in health and improvements in mortality over time, some regions have stagnated or even regressed including countries of the former Soviet union which have experienced setbacks in adult mortality–especially for males (Salomon and Murray, 2002; Lopez et al., 2006). The clustering method employed here is able to identify this pattern in the former Soviet states as well as in North America and Asia. Among the non-former Soviet countries that at some point experience this pattern, only Hungary, Poland, and the Maori of New Zealand have not transitioned to the "Post Transition" pattern. Interestingly, the family from which this pattern emerges contains a range of start dates that is exactly the same as those for the "Mid Transition" pattern (1945-2005), but the median start date is 20 years later (1965 for Mid Transition and 1985 for this pattern).

The pattern emerging from family 5 could be called the "Modified Mid Transition" because it closely mirrors the "Mid Transition" pattern except for elevated mortality in the adult years for males. Diseases of the cardiovascular, circulatory, and respiratory systems along with injuries and accidents can account for a substantial portion of the excess adult mortality in these eastern European countries (Guo, 1993; Notzon et al., 1998; McKee and Shkolnikov, 2001). Analysts who examine mortality in this part of the world often attribute these primary causes of death to a series of indirect origins including alcohol consumption and smoking (Bobak and Marmot, 1996; McKee and Shkolnikov, 2001), inadequate health care (Andreev et al., 2003), and "social stress" associated with abrupt economic shifts (Shkolnikov and Cornia, 1998).

3.4.3 Family-specific Parameters

Tables 3.4, 3.5, 3.6, 3.7, 3.8 and 3.9 in Appendix C contain the component score vectors, median coefficients, and the below- and above-median deviations for males and females

that when inserted into Equations 3.7 through 3.9 produce the model life table schedules.

3.4.4 Evaluating System Performance

I assess the performance of the system in two ways, starting with how well it can classify an arbitrary mortality indicator into one of the families. Because users typically approach model life table systems with at least some information on the age pattern of mortality (e.g. ${}_1m_0$, ${}_5q_0$, ${}_{45}q_{15}$), I test the use of discriminant analysis to classify partial mortality schedules. Using the known classification from the cluster analysis as a training data set for a DA model, I attempt to reclassify each schedule from the HMD using one of four child mortality indicators alone and combining them with ${}_{45}q_{15}$ as a second mortality measure to determine the appropriate family. I compare the known classification from the cluster analysis of the complete schedules with the classification of the mortality indicators resulting from discriminant analysis. The percentage of HMD schedules that were misclassified are presented in table 3.1. For all four child mortality measures alone, roughly 25-29 percent of schedules were misclassified, while the addition of ${}_{45}q_{15}$ reduces misclassification by roughly 10-13 percentage points across all measures. When an observation is misclassified, it is typically misclassified into an adjacent cluster because when clusters are close or overlapping classification uncertainty is greatest.

Table 3.1: Percentage of HMD schedules misclassified when using one of four child mortality measures alone and combined with ${}_{45}q_{15}$ to train a Discriminant Analysis model.

	without ${}_{45}q_{15}$		**with** ${}_{45}q_{15}$	
	Male	Female	Male	Female
${}_1m_0$	28.8	28.1	15.2	14.8
${}_5m_0$	26.9	25.5	14.8	14.2
${}_1q_0$	28.9	28.2	14.9	15.9
${}_5q_0$	26.5	25.4	14.5	14.3

I now turn to assessing the ability of this model to reproduce age-specific mortality rate schedules, first with an in-sample validation where I fit the 844 mortality rate schedules from the HMD with the model described above. I first identify the appropriate family using discriminant analysis to find the family that most closely matches ${}_5q_0$ (probability

of death before reaching age 5) and ${}_{45}q_{15}$ (the probability of death between age 15 and 60). I then select the level by finding the α value that creates a schedule that matches either life expectancy at birth, ${}_5q_0$, or ${}_{45}q_{15}$. I report the mean absolute error among all life tables and age groups as well as for e_0, ${}_5q_0$, and ${}_{45}q_{15}$. For comparison to existing model life table systems I compare the fit from this system to the Wilmoth model (Wilmoth et al., 2012), the WHO modified logit model (Murray et al., 2003), Coale and Demeny model life tables (Coale and Demeny, 1966), and the UN model life tables for developing countries (United Nations, 1982).[4]

Results are presented in table 3.2. For all of the validation metrics, smaller numbers indicate better fit and the lowest value in each column is bolded. For the HMD calibrated model, using life expectancy at birth produces the best overall fits which is to be expected given that it uses the most input information of the three combinations. Overall the HMD calibrated model performs similarly to that of the WHO and Wilmoth models. The Wilmoth model having also been calibrated with the HMD performs best on the mean absolute error amongst all age groups and has the lowest MAE for life expectancy as well. The Wilmoth model's slightly better fit in terms of the All-ages MAE may be attributed to the fact that his model has two continuous parameters rather than one discrete and one continuous as is the case with the family/level setup seen in the Coale and Demeny and UN model life tables as well as the HMD calibrated system described in this chapter. Overall the WHO modified logit, Wilmoth and HMD calibrated model life table system presented here achieve similar fits with just 0.00048 (in original scale) separating the high and low All-ages MAE amongst all three input combinations for the HMD model, the WHO, and Wimoth models for females and 0.0016 for males.[5] All three of these models are able to produce more accurate estimates of age-specific mortality rates and mortality indicators than those produced by the Coale and Demeny and UN model life tables.

---

[4]The UN system is fit using the procedure described on page 2 of Chapter IV of United Nations (1982), *Model Life Tables for Developing Countries* where I use the complete set of age-specific probabilities of death to estimate the appropriate loading factor. To fit the Coale and Demeny life tables I select the level by matching life expectancy to the closest half year and select the region that minimizes the sum of squared errors from the observed mortality rates. The WHO system was fit using the STATA software with the function modmatch (available at: `http://www.who.int/healthinfo/global_burden_disease/tools_software/en/`). The Wilmoth et al. model was fit with code provided at `http://www.demog.berkeley.edu/~jrw/LogQuad/`.

[5]Wilmoth et al. also find negligible differences in error between the performance of their model and that of a re-calibrated (with HMD data) WHO modified logic method (Wilmoth et al., 2012, p. 23).

Table 3.2: Validation results: Mean absolute error for three mortality indicators and amongst all-ages and life tables after fitting HMD life tables with various all-age mortality models.[†]

Model	Male				Female			
	All-ages[‡]	e_0	${}_5q_0$	${}_{45}q_{15}$	All-ages[‡]	e_0	${}_5q_0$	${}_{45}q_{15}$
HIV MLT								
input: e_0	2.61	—	**9.77**	**20.27**	1.76	—	**6.83**	**13.14**
input: ${}_5q_0$	3.01	1.59	—	36.75	2.01	1.26	—	22.24
input: ${}_5q_0$ and ${}_{45}q_{15}$	3.30	1.85	23.69	—	2.18	1.39	15.23	—
UN	4.59	1.05	—	—	3.07	2.09	—	—
CD[§]	5.09	—	34.36	52.13	3.70	—	28.51	39.71
WHO[§]	2.98	0.80	—	—	2.08	0.71	—	—
Wilmoth[§]	**2.14**	**0.46**	—	—	**1.70**	**0.53**	—	—

[†] The mean absolute errors for 'All-ages', ${}_5q_0$, and ${}_{45}q_{15}$ are expressed per 1,000

[‡] 'All-ages' refers to the mean absolute error for the non-logged mortality rates across age groups (0, 1-4, 5-9, 10-14, ..., 75) and amongst all life tables ($1, ..., N = 844$).

[§] 'WHO,' 'CD,' and 'Wilmoth' contain blank spaces as these quantities are inputs to these systems and thus have no error.

As one might expect, the model presented here along with the Wilmoth model perform well when fitting the data on which they were calibrated. To further test this method for model life table construction and compare it to existing systems, I also perform a cross validation where I calibrate a model with a random 75% sample and then fit the remaining 25%. I also fit the remaining 25% sample with the three existing model life table systems. This fitting procedure is performed iteratively 1,000 times yielding a distribution of 1,000 tables similar to table 3.2. Table 3.3 reports the mean errors from these distributions. Again, smaller numbers indicate better fit and the smallest number in each column is bolded.

Results from the iterative cross validation confirm the in-sample validation results. The method for life table construction presented in this paper performs similarly to the Wilmoth and WHO models. This result is not surprising as as those models were calibrated with large, contemporary datasets containing a wide range of mortality experiences similar to the HMD calibrated model. Again, the new method here shows small differences

in errors for the three most recent systems, which all outpace the Coale and Demeny and UN tables.

Table 3.3: Mean of the distribution of mean absolute error for three mortality indicators and amongst all-ages and life tables after fitting HMD life tables with various all-age mortality models after 1,000 iterations of cross validation.[†]

Model	Male				Female			
	All-ages[‡]	e_0	${}_5q_0$	${}_{45}q_{15}$	All-ages[‡]	e_0	${}_5q_0$	${}_{45}q_{15}$
HIV MLT								
input: e_0	2.66	—	**9.47**	**20.55**	1.93	—	**6.97**	**14.58**
input: ${}_5q_0$	3.00	1.64	—	37.20	2.09	1.33	—	23.32
input: ${}_5q_0$ and ${}_{45}q_{15}$	3.33	1.84	22.68	—	2.24	1.40	14.64	—
UN	4.60	1.05	12.02	24.72	3.07	2.09	19.67	17.39
CD[§]	5.09	—	34.36	52.11	3.70	—	28.49	39.67
WHO[§]	2.99	0.80	—	—	2.08	0.71	—	—
Wilmoth[§]	**2.14**	**0.46**	—	—	**1.70**	**0.53**	—	—

[†] The mean absolute errors for 'All-ages', ${}_5q_0$, and ${}_{45}q_{15}$ are expressed per 1,000

[‡] 'All-ages' refers to the mean absolute error for the non-logged mortality rates across age groups (0, 1-4, 5-9, 10-14, ..., 75) and amongst all life tables ($1, ..., L = 312$).

[§] 'WHO,' 'CD,' and 'Wilmoth' contain blank spaces as these quantities are inputs to these systems and thus have no error.

The near indistinguishable fits among the WHO, Wilmoth and the model under discussion here are demonstrated in figure 3.9, which shows the predicted fits from all four comparison models and for all three input combinations for the HMD calibrated model for US Males for the period 2000-2004. This figure, which depicts a typical fit among the 844 schedules fit for validation, shows that the three most recent systems all produce mortality rates that fall very closely to the data in nearly all cases. This is a very recent set of mortality rates and it is these more recent schedules (this one with low childhood mortality) that generate the largest errors for the Coale and Demeny and UN systems.

3.5 Discussion

This chapter presents a parsimonious, robust model of age-specific mortality and a method for identifying common age patterns of mortality. Together these allow for the creation of a traditional system of model life tables that includes a small number of families with

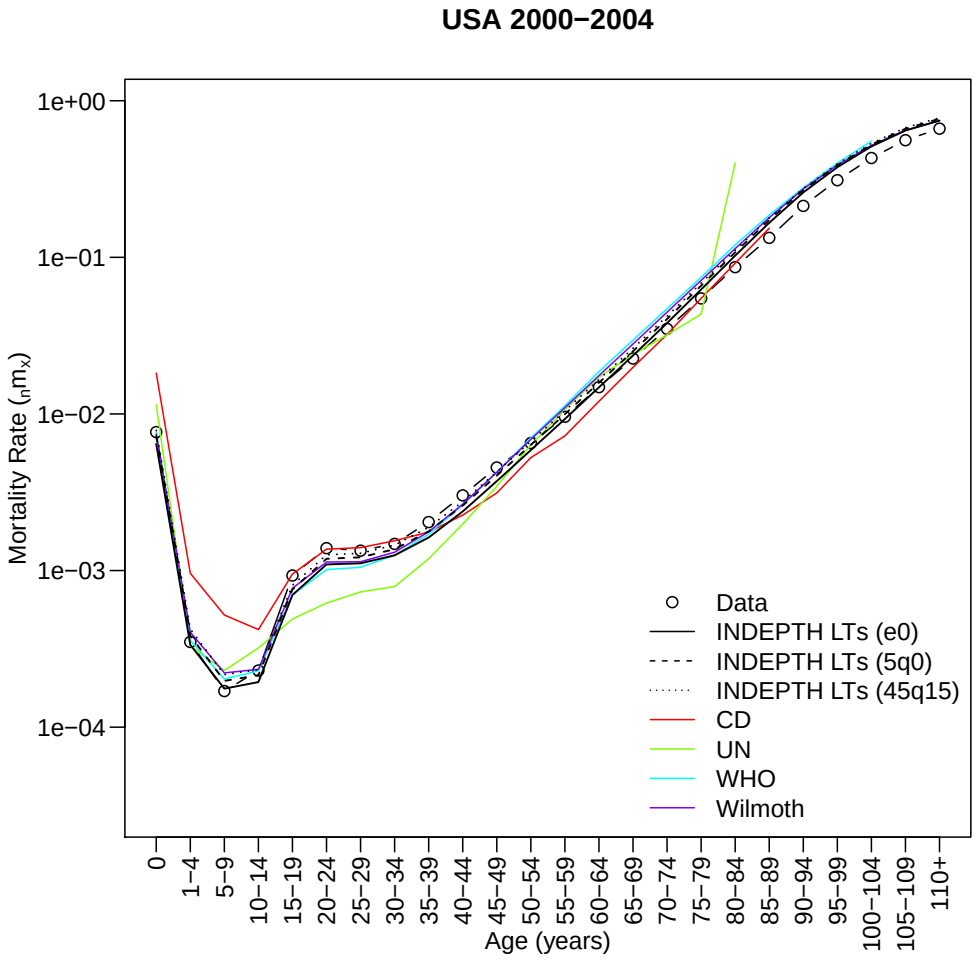


Figure 3.9: Fits of HMD model with three different input combinations to determine the level for USA Males 2000-2004. 1.) Life expectancy at birth [solid black line] 2.) Child mortality [dashed black line] 3.) Adult mortality [dotted black line]. For comparison fits from the WHO modified logit model [teal solid line], Coale and Demeney model life tables [red solid line], UN model life tables for developing countries [green solid line], and the Wilmoth et al. model [purple solid line] are also shown.

varying levels of mortality within each family. I 'calibrate' this model using the life tables in the HMD life table database to produce a model life table system that describes the variation in age-patterns of mortality in the HMD life tables. The clustering method, which allows similar patterns to reveal themselves, also allows one to draw substantive conclusions concerning the temporal and regional nature of mortality.

To validate this system, I present various measures of goodness of fit and conduct both in- and out-of-sample validation. The new model life table system performs well compared to several existing and widely used model life table systems. The new families identified in the cluster step include some with low child mortality and relatively high adult mortality that do not appear in any existing model life table systems – i.e. families 3 and 4.

Missing from the calibrated model is mortality experience in the developing world or in populations with unusual demographic or epidemiological circumstances, e.g. war-time mortality or high HIV prevalence. I address this limitation through a collaboration with the INDEPTH network (INDEPTH, 2011) of health and demographic surveillance system sites in Africa and Asia in the following chapter. INDEPTH has compiled a large collection of life tables from its member sites, and in the next chapter I apply this model life table framework to the INDEPTH collection.

Another drawback of this method for model life table construction is that, unlike the Wilmoth model, this model maintains the one discrete and one continuous parameter configuration of the CD and UN models. Although we show similar fits, the Wilmoth model does slightly better and the model presented could be improved by moving toward a more continuous configuration. Being able to move between family patterns would approximate the continuous setup and possibly lead to even better fits. This combination of patterns is possible by weighting the patterns according to the likelihoods of falling into a given classification (family) produced by the discriminant analysis step used to select the appropriate family pattern. Assuming one has the requisite pieces of information to use as input parameters, choosing between these two models is ultimately an arbitrary choice–both work best with two input parameters, but both will work with just one piece of information as well, both are calibrated with a large, diverse dataset with sufficient variation, and both determine the shape and level of mortality with each of the input

parameters but the flexibility of the method presented here may be slightly greater. For example, the Wilmoth model in the two input parameter case does require information about child *and* adult mortality–i.e. you cannot supply ${}_1q_0$ and ${}_5q_0$ because there is no direct information on adult mortality–where as the method presented here does not have any restrictions on the age-specific mortality rates or indicators one can use to produce a predicted set of mortality rates. For the method presented here, a single indicator such as ${}_5q_0$ can be used to determine both the shape and level of mortality even without additional information on adult mortality.

The model life table system presented in this chapter maintains the attractive features of the traditional model life table systems – i.e. parsimony, intuitive interpretation andease of use – while introducing a new fully automated and statistically valid method of construction. This removes any 'analyst bias' from the process of identifying the empirical regularities in a collection of life tables and produces reproducible classifications of life tables to families. This model of age-specific mortality allows easy, flexible manipulation of mortality age-patterns which enables one to construct a simple model of mortality age-patterns within each family. Perhaps most important, this framework (along with the tool described immediately below) is easy to use and relatively flexible. The user can approach the model life table system with a variety of summary mortality indicators with which one can identify the most similar mortality family and appropriate level within that family. The discriminant analysis functionality is flexible and allows the user to train the system using an arbitrary combination of mortality indicators not necessarily included in the basic release of the tool (below). Finally, this general approach (essentially a linear system) easily lends itself to many interesting and useful additions such as interpolating between two or more families.

To make this model life table system accessible and useful, I have created a package with a user-friendly graphical interface that implements the HMD-calibrated version of this model for the free, open source statistical package `R` (R Development Core Team, 2011) called `LifeTables`. The package is available from the `CRAN` archive or can be requested from the author.

Appendix A - Fixed model parameters

Table 3.4: Component score vector values from HMD calibration (left-singular vectors from SVD of HMD data): **S**. When inserted into equation 3.1 with **B** from table 3.5 produces the underlying family patterns.

	Male				Female			
Age	v1	v2	v3	v4	v1	v2	v3	v4
0	-0.10942	-0.36456	0.31976	0.12879	-0.11662	-0.35830	0.30010	0.12385
1-4	-0.19268	-0.37429	0.03399	0.05867	-0.19766	-0.40410	0.02355	0.04919
5-9	-0.22114	-0.13449	0.02214	0.00020	-0.22922	-0.19411	-0.05173	-0.05698
10-14	-0.22648	0.00536	-0.03961	-0.04350	-0.23536	-0.09290	-0.12990	-0.12124
15-19	-0.20065	0.11199	-0.16963	-0.10127	-0.21898	-0.04539	-0.20526	-0.17254
20-24	-0.18876	0.08125	-0.23927	-0.08222	-0.21238	-0.07447	-0.19725	-0.14148
25-29	-0.18802	0.09426	-0.22089	0.01015	-0.20783	-0.07266	-0.14051	-0.09933
30-34	-0.18420	0.11459	-0.15548	0.07039	-0.20056	-0.03784	-0.09478	-0.05204
35-39	-0.17647	0.13395	-0.08794	0.13676	-0.19109	0.00947	-0.04720	0.01364
40-44	-0.16598	0.15950	-0.01295	0.15460	-0.18068	0.06429	-0.00251	0.03775
45-49	-0.15401	0.18246	0.04951	0.16376	-0.16953	0.11633	0.04152	0.06177
50-54	-0.14137	0.19088	0.10486	0.15206	-0.15761	0.12349	0.07394	0.06957
55-59	-0.12877	0.19452	0.14800	0.13230	-0.14562	0.12054	0.10188	0.07713
60-64	-0.11538	0.17457	0.16354	0.12220	-0.13168	0.08816	0.13586	0.08019
65-69	-0.10226	0.15357	0.16968	0.08376	-0.11710	0.06797	0.17252	0.04367
70-74	-0.08846	0.12047	0.16837	0.04343	-0.10088	0.04237	0.19804	0.01891
75-79	-0.07464	0.09262	0.15874	0.00487	-0.08460	0.03105	0.20083	-0.00184
80-84	-0.06118	0.07059	0.14972	-0.07238	-0.06893	0.03329	0.19223	-0.06298
85-89	-0.04834	0.05471	0.14185	-0.14277	-0.05434	0.03620	0.17129	-0.13905
90-94	-0.03706	0.04997	0.12906	-0.21531	-0.04167	0.04102	0.14671	-0.20837
95-99	-0.02727	0.03711	0.11264	-0.24883	-0.03042	0.04044	0.12333	-0.25367
100-104	-0.01941	0.03115	0.09599	-0.26567	-0.02138	0.04149	0.09961	-0.27718
105-109	-0.01337	0.02670	0.07979	-0.26262	-0.01446	0.04043	0.07889	-0.27997
110+	-0.00954	0.02356	0.06866	-0.25294	-0.01012	0.03806	0.06600	-0.27463

Table 3.5: Median coefficients for first 4 score vectors for HMD calibration, by cluster/family, **B**. When inserted into equation 3.1 with **S** from table 3.4 produces the underlying family patterns.

Cluster	Intercept	Coeff 1	Coeff 2	Coeff 3	Coeff 4
1	0.0500	29.0949	-1.2763	-0.0014	-0.0681
2	0.0269	33.5397	1.5962	-0.2199	0.6667
3	-0.0381	34.0455	0.6514	0.8543	-0.1176
4	0.0714	37.9729	1.6700	-0.7928	-0.1456
5	-0.0052	24.2750	-2.7641	-0.1395	-0.0118

Table 3.6: Male above-median family-specific deviations and cluster-invariant deviations for HMD calibration: $\mathbf{D}_{f+}$

	Family					
Age	1	2	3	4	5	$\mathbf{D}_{h+}$
0	0.737	1.015	0.912	0.572	0.558	2.277
1-4	1.172	1.181	1.051	0.684	0.828	3.178
5-9	0.754	0.913	0.737	0.680	0.545	2.590
10-14	0.620	0.772	0.555	0.626	0.485	2.207
15-19	0.683	0.662	0.400	0.473	0.603	1.746
20-24	0.901	0.754	0.465	0.425	0.803	1.835
25-29	0.955	0.860	0.504	0.431	0.916	1.837
30-34	0.868	0.880	0.494	0.443	0.889	1.722
35-39	0.788	0.905	0.484	0.446	0.876	1.583
40-44	0.630	0.849	0.428	0.440	0.769	1.359
45-49	0.492	0.790	0.373	0.425	0.675	1.139
50-54	0.365	0.719	0.324	0.409	0.570	0.951
55-59	0.260	0.649	0.279	0.390	0.478	0.786
60-64	0.258	0.632	0.280	0.366	0.469	0.725
65-69	0.242	0.582	0.266	0.337	0.430	0.657
70-74	0.257	0.545	0.269	0.308	0.411	0.629
75-79	0.284	0.511	0.269	0.274	0.407	0.594
80-84	0.255	0.417	0.235	0.237	0.333	0.513
85-89	0.222	0.327	0.198	0.201	0.264	0.421
90-94	0.179	0.228	0.148	0.163	0.189	0.309
95-99	0.200	0.196	0.138	0.133	0.189	0.266
100-104	0.229	0.183	0.133	0.106	0.209	0.228
105-109	0.278	0.197	0.141	0.086	0.258	0.214
110+	0.319	0.218	0.151	0.073	0.303	0.213

Table 3.7: Male below-median family-specific deviations and cluster-invariant deviations for HMD calibration: $\mathbf{D}_{f-}$

	Family					
Age	1	2	3	4	5	$\mathbf{D}_{h-}$
0	0.731	0.902	0.911	0.777	1.069	1.934
1-4	1.061	1.168	0.986	0.868	1.187	2.245
5-9	0.791	0.960	0.688	0.709	0.844	1.779
10-14	0.697	0.859	0.520	0.614	0.691	1.466
15-19	0.717	0.799	0.393	0.521	0.661	1.110
20-24	0.850	0.858	0.456	0.562	0.755	1.157
25-29	0.874	0.832	0.489	0.620	0.707	1.192
30-34	0.809	0.769	0.481	0.635	0.630	1.169
35-39	0.746	0.700	0.477	0.651	0.555	1.139
40-44	0.636	0.619	0.433	0.623	0.481	1.048
45-49	0.538	0.546	0.391	0.591	0.421	0.952
50-54	0.451	0.492	0.355	0.553	0.391	0.871
55-59	0.379	0.450	0.323	0.514	0.377	0.797
60-64	0.372	0.447	0.332	0.503	0.411	0.782
65-69	0.361	0.452	0.328	0.474	0.458	0.751
70-74	0.370	0.472	0.339	0.451	0.528	0.744
75-79	0.386	0.492	0.347	0.429	0.595	0.727
80-84	0.369	0.506	0.325	0.374	0.661	0.670
85-89	0.349	0.514	0.301	0.321	0.716	0.607
90-94	0.322	0.515	0.264	0.262	0.758	0.524
95-99	0.334	0.527	0.261	0.240	0.803	0.494
100-104	0.351	0.535	0.261	0.228	0.833	0.468
105-109	0.378	0.543	0.271	0.232	0.854	0.460
110+	0.402	0.548	0.282	0.242	0.868	0.461

Table 3.8: Female above-median family-specific deviations and cluster-invariant deviations for HMD calibration: $\mathbf{D}_{f+}$

	Family					
Age	1	2	3	4	5	$\mathbf{D}_{h+}$
0	0.755	1.021	0.914	0.581	0.569	2.328
1-4	1.221	1.205	1.091	0.702	0.843	3.316
5-9	0.915	0.942	0.806	0.688	0.611	2.875
10-14	0.845	0.818	0.666	0.641	0.571	2.624
15-19	0.883	0.740	0.581	0.552	0.621	2.379
20-24	0.951	0.800	0.631	0.548	0.689	2.437
25-29	0.888	0.816	0.633	0.559	0.661	2.372
30-34	0.807	0.815	0.599	0.549	0.646	2.193
35-39	0.728	0.827	0.558	0.531	0.653	1.969
40-44	0.596	0.773	0.480	0.505	0.593	1.685
45-49	0.469	0.722	0.405	0.478	0.537	1.406
50-54	0.413	0.696	0.379	0.457	0.506	1.265
55-59	0.381	0.681	0.367	0.437	0.490	1.160
60-64	0.375	0.681	0.386	0.422	0.478	1.126
65-69	0.300	0.611	0.361	0.401	0.389	1.017
70-74	0.270	0.570	0.353	0.374	0.346	0.931
75-79	0.271	0.536	0.337	0.334	0.342	0.824
80-84	0.227	0.441	0.281	0.285	0.282	0.659
85-89	0.188	0.334	0.219	0.233	0.217	0.502
90-94	0.160	0.238	0.163	0.185	0.166	0.361
95-99	0.165	0.181	0.130	0.144	0.154	0.266
100-104	0.192	0.156	0.112	0.109	0.173	0.198
105-109	0.238	0.162	0.113	0.083	0.220	0.166
110+	0.278	0.177	0.120	0.068	0.261	0.157

Table 3.9: Female below-median Family-Specific deviations and cluster-invariant deviations for HMD calibration: $\mathbf{D}_{f-}$

	Family					
Age	1	2	3	4	5	$\mathbf{D}_{h-}$
0	0.746	0.914	0.908	0.780	1.066	1.945
1-4	1.099	1.210	1.020	0.884	1.232	2.323
5-9	0.909	1.073	0.746	0.725	0.976	1.907
10-14	0.863	1.023	0.616	0.642	0.890	1.669
15-19	0.877	1.001	0.548	0.581	0.898	1.476
20-24	0.917	1.019	0.595	0.614	0.934	1.538
25-29	0.870	0.978	0.600	0.628	0.892	1.541
30-34	0.807	0.908	0.573	0.626	0.813	1.467
35-39	0.741	0.821	0.540	0.630	0.714	1.370
40-44	0.643	0.727	0.476	0.597	0.616	1.224
45-49	0.549	0.636	0.414	0.565	0.524	1.080
50-54	0.505	0.597	0.397	0.549	0.504	1.022
55-59	0.477	0.571	0.394	0.540	0.502	0.987
60-64	0.469	0.570	0.419	0.541	0.540	1.005
65-69	0.419	0.554	0.406	0.504	0.569	0.964
70-74	0.396	0.548	0.409	0.480	0.610	0.940
75-79	0.390	0.541	0.405	0.457	0.646	0.889
80-84	0.359	0.528	0.364	0.398	0.677	0.783
85-89	0.332	0.522	0.318	0.331	0.717	0.669
90-94	0.313	0.519	0.275	0.271	0.753	0.564
95-99	0.314	0.521	0.252	0.234	0.789	0.493
100-104	0.328	0.525	0.242	0.214	0.817	0.443
105-109	0.354	0.530	0.245	0.212	0.838	0.421
110+	0.376	0.536	0.253	0.218	0.853	0.416

Table 3.10: Alpha values to index family-specific 'levels' by life expectancy at birth for HMD calibration, males

	Family				
Target e_0	1	2	3	4	5
30	0.982	1.477	1.505	1.911	0.487
32.5	0.934	1.425	1.461	1.868	0.423
35	0.884	1.371	1.416	1.823	0.352
37.5	0.831	1.314	1.369	1.776	0.272
40	0.774	1.253	1.318	1.727	0.177
42.5	0.713	1.187	1.265	1.675	0.060
45	0.645	1.116	1.208	1.619	-0.070
47.5	0.571	1.039	1.145	1.559	-0.182
50	0.487	0.957	1.076	1.494	-0.290
52.5	0.397	0.867	0.999	1.429	-0.393
55	0.287	0.764	0.920	1.355	-0.495
57.5	0.146	0.645	0.828	1.272	-0.603
60	-0.051	0.506	0.717	1.174	-0.714
62.5	-0.250	0.347	0.580	1.060	-0.828
65	-0.431	0.150	0.408	0.929	-0.947
67.5	-0.611	-0.100	0.162	0.772	-1.075
70	-0.798	-0.381	-0.218	0.565	-1.216
72.5	-0.988	-0.635	-0.525	0.281	-1.364
75	-1.201	-0.885	-0.816	-0.152	-1.526
77.5	-1.425	-1.146	-1.099	-0.506	-1.709
80	-1.667	-1.395	-1.390	-0.842	-1.904
82.5	-1.935	-1.665	-1.697	-1.176	-2.122
85	-2.221	-1.942	-2.005	-1.517	-2.361
87.5	-2.524	-2.236	-2.338	-1.877	-2.622
90	-2.851	-2.540	-2.684	-2.253	-2.906

Table 3.11: Alpha values to index family-specific 'levels' by life expectancy at birth for HMD calibration, females

	Family				
Target e_0	1	2	3	4	5
30	1.022	1.578	1.562	1.952	0.533
32.5	0.979	1.534	1.521	1.916	0.474
35	0.937	1.490	1.481	1.878	0.415
37.5	0.892	1.447	1.441	1.840	0.349
40	0.845	1.401	1.399	1.800	0.273
42.5	0.796	1.354	1.355	1.758	0.183
45	0.742	1.305	1.309	1.715	0.067
47.5	0.684	1.249	1.259	1.668	-0.062
50	0.620	1.190	1.205	1.619	-0.167
52.5	0.549	1.126	1.147	1.565	-0.267
55	0.470	1.056	1.082	1.507	-0.364
57.5	0.383	0.979	1.010	1.448	-0.460
60	0.276	0.897	0.935	1.383	-0.558
62.5	0.130	0.802	0.849	1.308	-0.661
65	-0.078	0.690	0.746	1.223	-0.767
67.5	-0.275	0.557	0.619	1.121	-0.876
70	-0.455	0.399	0.460	1.003	-0.991
72.5	-0.636	0.196	0.249	0.867	-1.119
75	-0.824	-0.087	-0.120	0.691	-1.255
77.5	-1.016	-0.371	-0.452	0.451	-1.402
80	-1.232	-0.638	-0.752	0.074	-1.565
82.5	-1.458	-0.904	-1.044	-0.389	-1.747
85	-1.715	-1.181	-1.355	-0.766	-1.949
87.5	-1.991	-1.472	-1.679	-1.146	-2.177
90	-2.303	-1.788	-2.028	-1.520	-2.434

Chapter 4

INDEPTH MODEL LIFE TABLES FOR LOW- AND MIDDLE-INCOME COUNTRIES

4.1 Introduction

In this chapter, I present a new model life table system for low- and middle-income countries calibrated with data from the INDEPTH life table database and fit using the method described in the preceding chapter. As noted in the introduction to this dissertation, roughly half of the world's countries do not have even a partial vital registration system that could give accurate and reliable estimates of age-specific mortality. Without data or estimates of the mortality profile, governments and non-governmental organizations alike may find it difficult to plan for the future or allocate already limited resources. In absence of this information, countries must rely on *model life tables* to estimate complete age-patterns of mortality from age-restricted mortality indicators such as child (${}_5q_0$) or adult (${}_{45}q_{15}$) mortality.

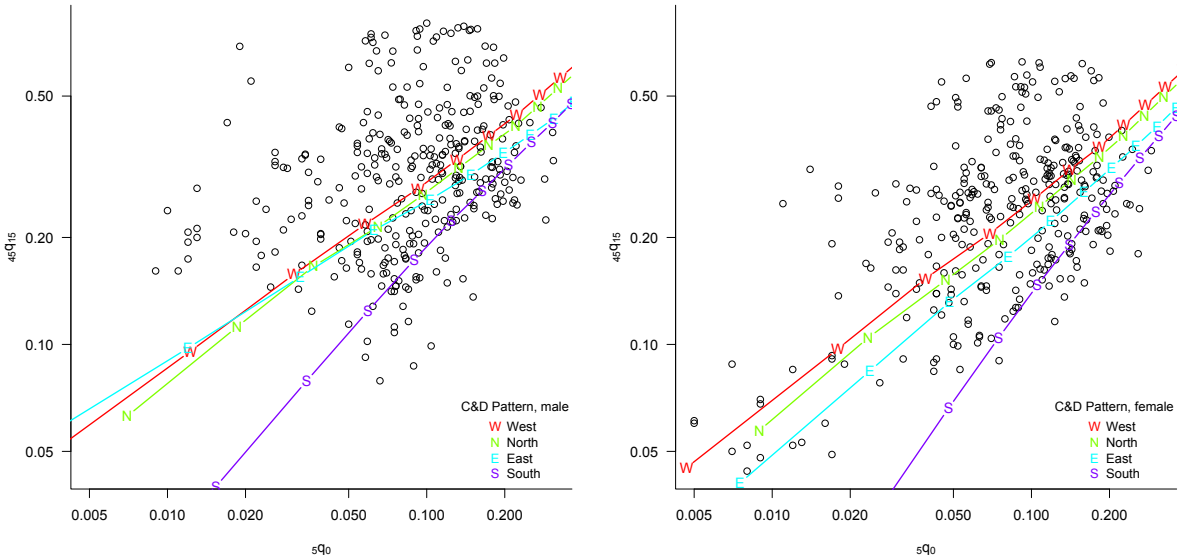
Current model life table systems - even those designed for use in developing settings - lack sufficient data from Africa and are based on data that is out dated (Coale and Demeny, 1966; Coale and Guo, 1989; United Nations, 1982; Murray et al., 2003; Wilmoth et al., 2012). The problem is compounded by the fact that in order to build a reliable model life table system one would need a fairly sizable collection of mortality data from the developing world where data are limited. INDEPTH's health and demographic surveillance sites (HDSS) represent one such source of data and an opportunity to build a model life table system for developing countries, especially in Africa, that reflects age patterns of mortality that other models cannot produce like those resulting from high HIV prevalence populations or those suffering from a high number of malaria deaths.

Although analysts still use existing systems to estimate mortality in Africa, there is little reason to expect contemporary African mortality profiles to reflect patterns that were generated under an entirely different set of social, epidemiological, and demographic conditions. Figure 4.1 plots the relationship between adult and child mortality for the

INDEPTH life table collection (each black circle represents one single-year life table) and the relationship between child and adult mortality at various levels in the regional patterns produced by the Coale and Demeny (1966) and United Nations (1982) systems. These figures show that mortality profiles with relatively low child mortality accompanied by elevated adult mortality are not well represented by any of the patterns in these two systems. The cloud of black points in this figure also extends far above any of the available pattern and level combinations available in these two systems in the middle range of child mortality (center of horizontal axis). These life tables arise from high-HIV prevalence populations in southern Africa in the 1990s and 2000s. As with the similar figures presented in the preceding chapter, the inability of these models to represent modern mortality schedules is not the result of methodological shortcomings but rather is a reflection of the state of mortality at the time. For instance, high HIV-related mortality simply did not exist and is thus not represented in existing systems. Likewise, even more recent models like that of Wilmoth et al. (2012) are unable to produce patterns like those generated under high HIV prevalence, because despite having more recent data, it still lacks adequate information from Africa or other low-to-middle income countries. Figure 4.2 demonstrates that the Wilmoth model is unable to capture adult mortality from high HIV prevalence populations (black circles above the top dotted green line) or situations of relatively high child mortality likely resulting from a high Malaria related deaths among children (black circles below the bottom green dashed line).[1]

As with the HMD calibrated system described previously in this dissertation, I seek an easy-to-use model life table system modeled after several existing systems where 'families' characterize model mortality patterns with arbitrary 'levels' within each family. I identify commonly observed age-patterns of mortality in a collection of empirical life tables from the INDEPTH network of health and hemographic surveillance system sites, which represent populations throughout Africa and Southeast Asia including those experiencing high HIV prevalence (INDEPTH, 2011). Based on these commonly observed patterns, I build a model life table system and describe the temporal and regional patterns in mortality contained in the INDEPTH data.

[1]Wilmoth et al. do note that the 'k' parameter can take on values greater than 2 (or less than -2), but that in practice 'k' rarely exceeds an absolute value of 2 and that the patterns produced beyond this range become more implausible as you go beyond an absolute value of 4.



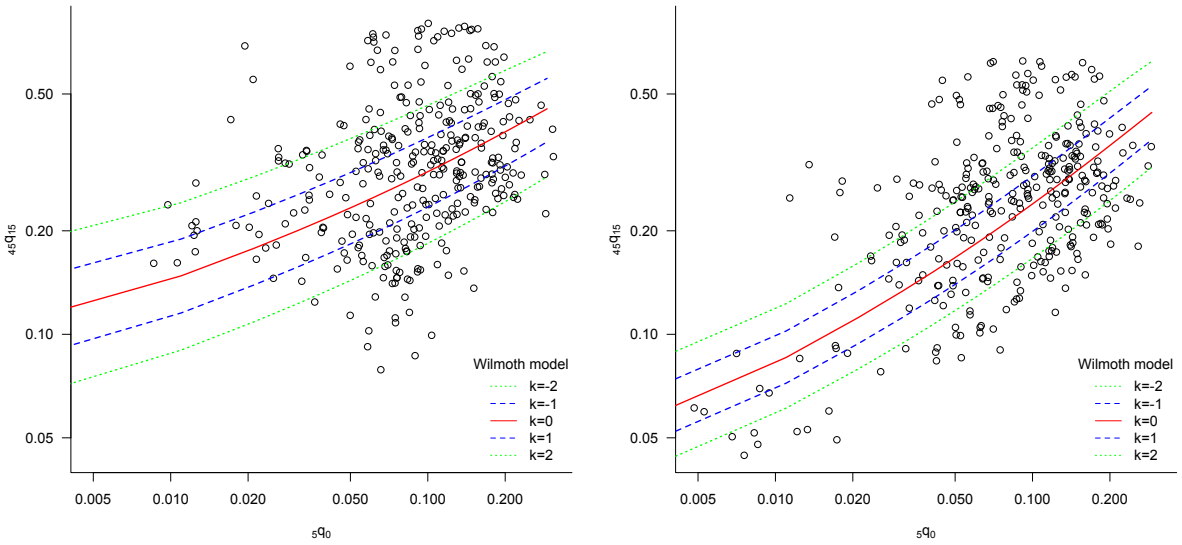
(a) C&D male

(b) C&D female

(c) UN male

(d) UN female

Figure 4.1: Relationship between child and adult mortality, observed INDEPTH data (n=312) and Coale-Demeny and UN regional model life table patterns



(a) Male

(b) Female

Figure 4.2: Relationship between child and adult mortality, observed INDEPTH data (n=329) and output from Wilmoth model resulting from varying the 'k' parameter. $_5q_0$ is the probability a newborn will die before reaching age 5 and $_{45}q_{15}$ is the probability a 15 year old will die before reaching age 60.

I identify typical age patterns by using a the model-based clustering method described in chapter three. This clustering method is effectively fully automated and permits the patterns to reveal themselves with very little input from the analyst. Using each cluster as the basis for a 'family' in a traditional system of model life tables, I generated levels of mortality within family with a one-parameter model. The result is a new, effectively two-parameter system of model life tables this time for the regions and time periods included in the INDEPTH data.

Below I provide a brief description of the INDEPTH data used to calibrate this system. I then present a brief description of the underlying model and resulting 'families'. The life table construction method employed in this chapter is described in detail in the previous chapter. Thus, I provide only a brief overview of the underlying model that produces a predicted set of mortality rates and describe a few minor methodological differences made necessary by the nature of this data. Finally, I discuss usage and validation of the system.

4.2 Data

I identify similar age patterns of mortality in a collection of life tables that covers low- and middle-income countries. This dataset contains 329 (for each sex) 1-year period life tables with age groups 0, 1-4, 5-9, ..., 85+ from the INDEPTH network. The mortality profiles in this collection represent 32 HDSS from Africa and southeast Asia with varying periods of coverage from 1983 to 2011.[2] Each HDSS longitudinally monitors defined subjects like individuals or households within a defined geographic area and contributes prospectively collected data on variable sized populations. All tables have been computed from directly observed deaths and person years.

The INDEPTH data were screened for consistency and plausibility before inclusion in the final data set. For each HDSS, I first plotted the raw data (age-specific deaths and person years) by year and removed HDSS-years that were out of trend or entire HDSS if the data looked implausible. I next calculated a life table for each HDSS-year and plotted the annual age-specific mortality rates as well as the annual trend in child mortality (${}_5q_0$), adult mortality (${}_{45}q_{15}$), and life expectancy at birth and removed any out-of-trend years. Data inclusion for each HDSS was also confirmed with a HDSS representative who could verify the trend in mortality data with information about epidemiological conditions at the HDSS or insights into data collection issues or protocol at the HDSS (e.g. child deaths may be systematically underreported as was the case for one Indian HDSS). Of the 37 HDSS that submitted mortality data, five were screened out entirely after an HDSS representative verified that data were incomplete or unreliable for all years in an HDSS. Of the remaining 32 HDSS, 12 had a portion of their data series removed. Typically, either the first year (or two) or final year of data from these HDSS was removed because the person years or deaths were lower than expected resulting from incomplete data collection in the most recent years or because suitable data collection protocols had not been solidified in the early years of an HDSS. After screening inapplicable data, the INDEPTH collection contains approximately 237,500 deaths and 23.2 million person years of observation.

---

[2]Figure 4.6 contains a listing of the regions providing input tables and the years of data availability.

4.3 Method

4.3.1 Model Life Table Construction

To construct the model life tables I follow the method detailed in chapter three of this dissertation. The model life table system presented in this chapter reflects the structure of previous systems with clusters of similar mortality rate schedules comprising "families" with varying "levels" of mortality within each family. A brief description of the underlying model that produces a complete set of mortality rates follows.

Assuming 19 age groups (0, 1-4, 5-9, 10-14,... 85+), a 19 x m matrix $\mathbf{M}$ composed of m column vectors of age-specific mortality rate schedules can be expressed as a weighted sum of a number of components whose shapes encode the fundamental age pattern of human mortality and a wide range of variations on that:

$$\mathbf{M} = \mathbf{SB} + \mathbf{C} + \mathbf{R} \tag{4.1}$$

$\mathbf{S}$ is a 19 x n matrix whose columns are the n 'components' used in the model. These 'components' are the left-singular vectors derived from a Singular Value Decomposition of all of the empirical mortality rate schedules and are plotted in figure 4.3. These components have the desirable property of encoding the bulk of the variation in the mortality rate schedule into a small number of vectors. $\mathbf{B}$ is a n x m matrix whose columns are coefficients that multiply each component schedule contained in $\mathbf{S}$ to yield the age-varying component of each mortality schedule. In order to obtain a family-specific set of coefficients, I first regress each empirical mortality rate schedule on the first few 'components' yielding a set of regression coefficients for each HDSS-year life table. I then use a model-based clustering method (Fraley and Raftery, 2006) to identify similar sets of coefficients. Each cluster gives rise to a 'family,' while the median set of coefficients in each cluster is the family-specific $\mathbf{B}$. $\mathbf{C}$ is a 19 x m matrix whose columns are constants that are added to modify each mortality schedule in an age-constant way generating 'levels' within family.[3] Finally, $\mathbf{R}$ is a 19 x m matrix of residuals that account for the remaining difference between the modeled and empirical mortality schedules. For each family, the pattern produced by the median set of coefficients without any adjustment to the level is

[3]A detailed description of the calculation of $\mathbf{C}$ can be found in section 3.3.3 of the previous chapter.

referred to as the 'underlying family pattern' (UFP), $\mathbf{M}_f$, where f indexes family.

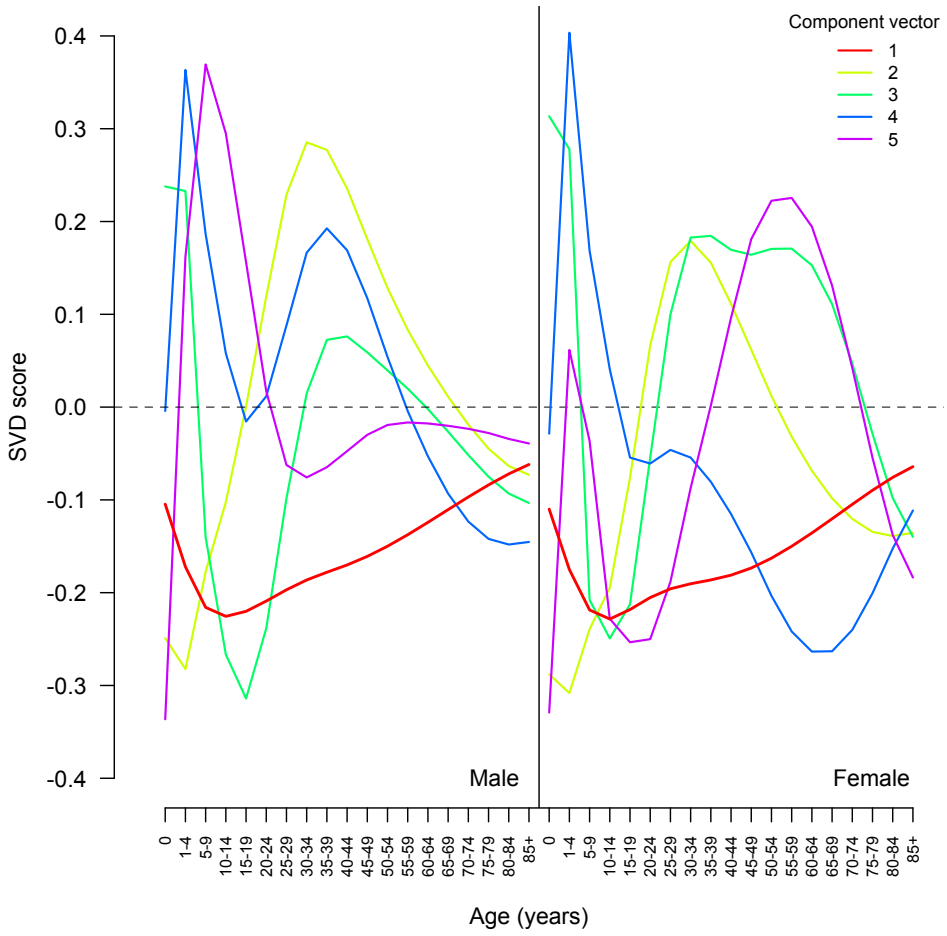


Figure 4.3: First five left singular vectors from the Singular Value Decomposition of the INDEPTH mortality rate schedules. **S** from equations 4.1 and 3.2.

Each mortality rate schedule naturally contains some stochastic variation producing meaningless bumps and wiggles in the schedules especially for HDSS with smaller populations (see figure 1.3). I use the eight parameter Heligman and Pollard (1980) described in chapter two to smooth any of this stochastic noise and perform the SVD on the smooth schedules.[4] I then regress the raw, un-smoothed mortality rate schedules on the left singu-

[4]The HP model captures the age-specific shape of mortality using three components: one for child mortality, one for old-age mortality and a necessarily flexible adult "hump" component, which works well to represent elevated mortality associated with high HIV-prevalence. However, the HP model is often unable to accurately reflect mortality under age 10 in the INDEPTH data. This issue arises because child mortality does not decline as precipitously as the model would imply. In order to maintain the shape of childhood mortality in these schedules, I combined the observed mortality rates for the first two age groups with the HP-smoothed schedules for the remaining age groups.

lar vectors obtained from the SVD. This procedure maintains the shape of the mortality rate schedules in each cluster while also producing a smooth underlying pattern for each family.

4.3.2 Usage

Users of the system can supply any of a variety of measures of child or adult mortality and discriminant analysis (DA) can be used to identify the family in which the supplied measure is most likely to fit. Once a training model is specified, DA can be used to 'predict' which classification or group an unknown case most likely belongs.[5] Once the appropriate family has been determined, the user can select a level of mortality by solving for the appropriate weight for $\mathbf{C}$ that produces a schedule with a desired life expectancy (or any other indicator).

4.4 Results

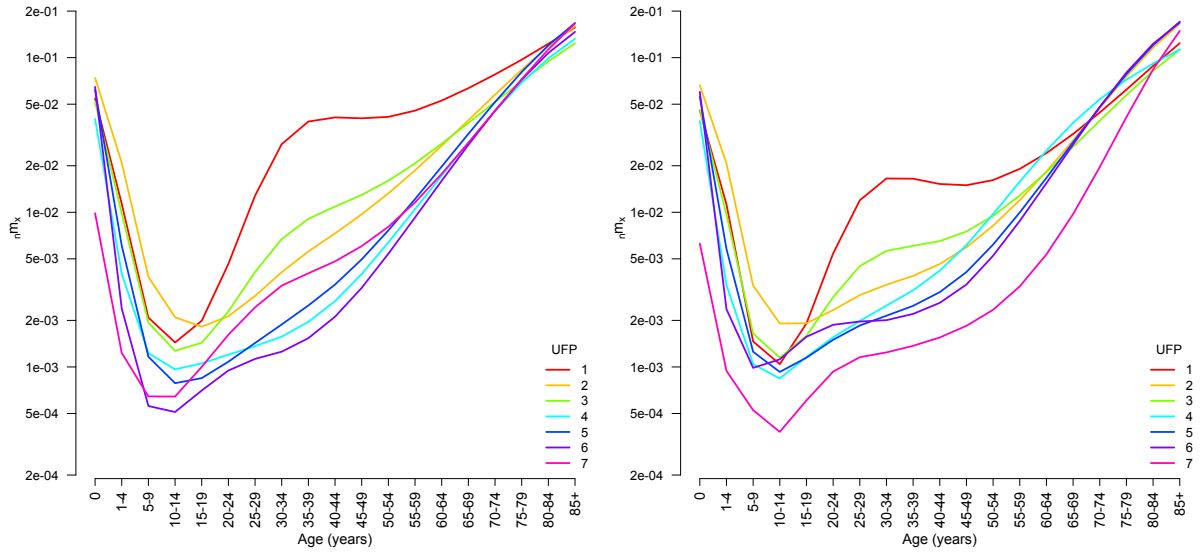
4.4.1 Mortality Patterns

Seven distinct mortality patterns emerge from the INDEPTH data. The UFPs, $\mathbf{M}_f$, are presented in figure 4.4. Each of these patterns arises from a family containing life tables from similar geographic areas or generated under similar epidemiological conditions.[6] Figure 4.5 plots the relationship between adult mortality (${}_{45}q_{15}$) and child mortality (${}_5q_0$) in the INDEPTH data (each small circle represents one single-year life table) as well as the resulting relationships from each model family after varying the level. As with the HMD figures from chapter three, one can see the INDEPTH patterns also cover a wide range of mortality experiences this time for low- and middle-income regions. For example, patterns 1 and 3 depict the high adult mortality humps associated with HIV while pattern 2 captures elevated child mortality likely resulting from malaria.

---

[5]The `R` package `Mclust`, which was used to perform the cluster analysis, contains functions to perform discriminant analysis where any portion of the age range or any summary measure like e_0 can be used with the known classification to train the DA model. These functions also provide a likelihood for classification into each cluster, which makes it possible to estimate a full set of mortality rates based on two or more clusters by weighting the underlying family patterns and deviations according to these likelihoods.

[6]Cause-specific data on deaths in each HDSS were unavailable thus I rely on national epidemiological conditions as a proxy.



(a) Male

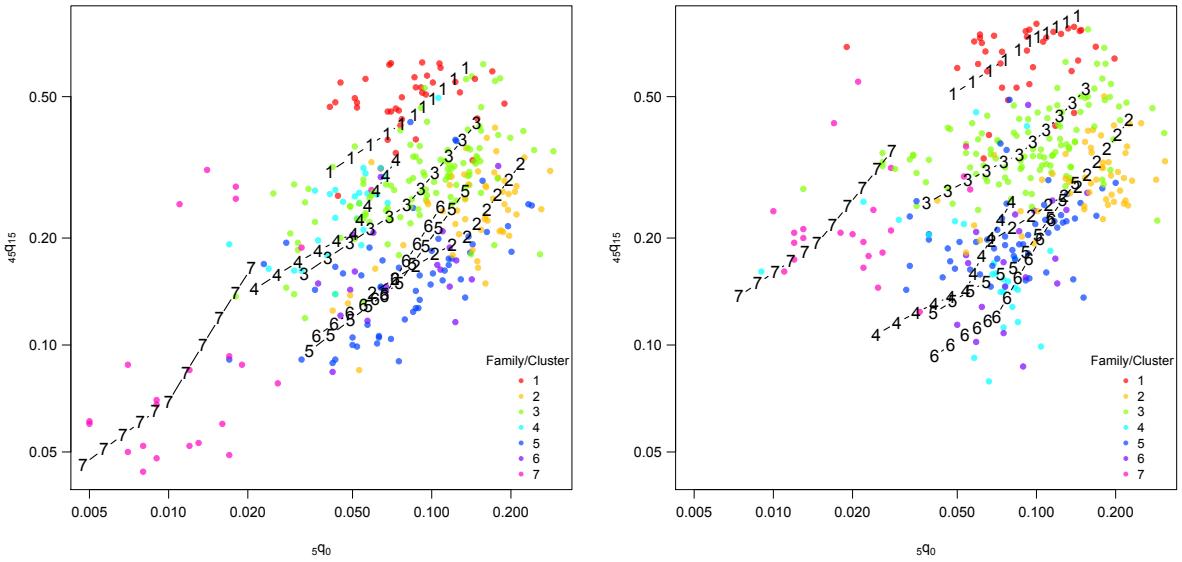
(b) Female

Figure 4.4: Underlying Family Age-Patterns of Mortality, $\mathbf{M}_f$, identified in the INDEPTH life table collection. y-axis shown in log scale

4.4.2 Components of each Cluster/Family

I now turn to a discussion of the cluster membership for each INDEPTH family touching on the differences between patterns in both form and cause of death. Figure 4.6 shows the distribution of cluster membership over time for each HDSS in the INDEPTH data set. Each circle, which is color coded according to the classification from the clustering step, represents one life table. I order the HDSS in this figure according their modal classification from the clustering step. The following description aggregates the patterns into four categories based on the geographic locations and epidemiological profiles that produce each pattern.

HIV Pattern Most life tables that make up pattern 1 originate from HDSS in South Africa, Kenya, and Mozambique. None of the tables in this family date prior to 2000 likely resulting from the timing of the HIV epidemic in these sites (UNAIDS, 2012; Department of Health, Republic of South Africa, 1995, 1996, 1997, 1998). Interestingly, the relatively longer data series for Agincourt reveals its transition from lower mortality (pattern 3) into



(a) Male

(b) Female

Figure 4.5: Child versus Adult Mortality in the INDEPTH life table collection and resulting relationships after varying α parameter (level). α varied from -1 to 1 with black numbers corresponding to family patterns at various levels. Observed data shown as small circles color coded according to cluster classification (family). Each circle represents one table from the INDEPTH collection.

the HIV pattern by 2002. This change in classification comes right around the time when AIDS-related mortality begins to substantially alter the mortality profile at Agincourt as described in chapter two. For both males and females, this pattern contains the highest mortality during the adult years and the second highest child mortality among all seven patterns. Although prominent in this life table collection, this pattern is relatively new and unique among existing model life table systems.

Malaria Pattern One could call pattern 2 the "Malaria" pattern as nearly all of the life tables in this family originate in west Africa where childhood mortality resulting from malaria is prevalent (World Health Organization, 2011; Murray et al., 2012) and HIV prevalence is relatively low compared to southern Africa. Adult mortality is lower than for those patterns influenced heavily by HIV but still generally higher than in patterns 4 through 7, which contain life tables exclusively from Asian HDSS.

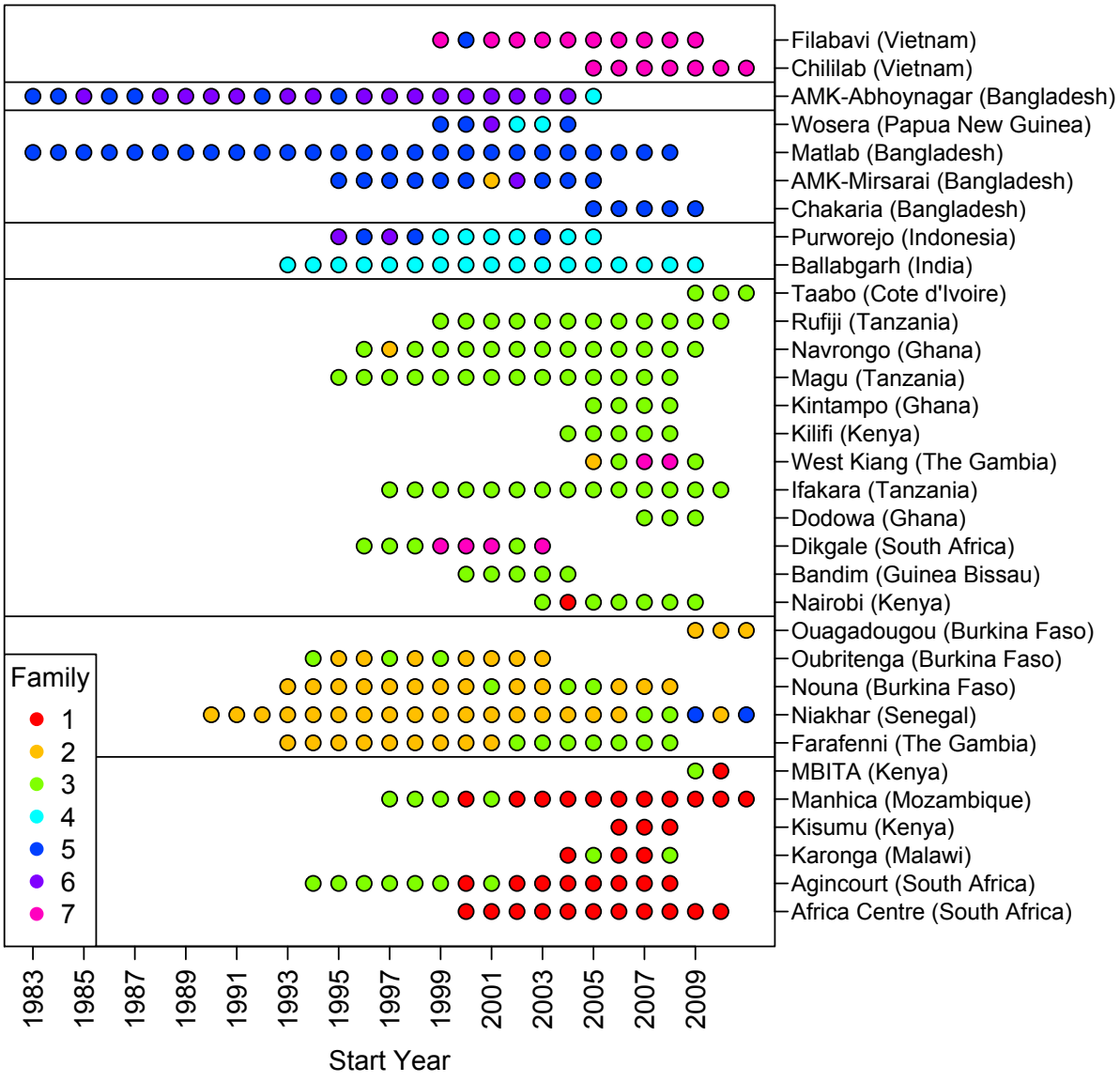


Figure 4.6: HDSS by Year and Family membership (cluster classification) for INDEPTH data. The HDSS are ordered along the vertical axis accruing to their model classification. The horizontal lines demarcate groups of HDSS with the same model classification.

HIV/Malaria Pattern Pattern 3 shows lower levels of adult mortality relative to pattern 1 with high mortality in the childhood years. Although lower than pattern 1, adult mortality is high relative to other families likely as a result of smaller HIV epidemics in some of these HDSS. The high under-5 mortality is possibly reflecting malaria deaths among children. Most life tables in this family originate from three HDSS in Tanzania but HDSS in west Africa, Kenya, and South Africa are also represented in this family.

Asian Patterns Life tables from Asian HDSS are concentrated in families 4 through 7 and appear to cluster along a geographic dimension with HDSS from India, Bangladesh, and Vietnam mainly comprising the respective families.

Pattern 4 arises from a family mostly containing life tables from just one HDSS in India with a few from Purworejo HDSS in Indonesia. For females, this pattern shows relatively low child mortality with adult mortality in the middle range for this data. For males, the low childhood mortality is also present but adult mortality is the second lowest only to pattern 6.

Life tables from HDSS in Bangladesh constitute the bulk of families 5 and 6. Among the Asian families, these patterns represent the highest level of under-5 mortality resulting from a number of causes including diarrheal disease and pneumonia (Black et al., 2010), while adult mortality is amongst the lowest for all families. Finally, family 7 is notable for its very low rates of childhood mortality–the lowest among all INDEPTH life tables. For females this pattern also represents the lowest adult mortality and shows the highest life expectancy of any of underlying INDEPTH patterns. For males in family 7, the low childhood mortality is followed by an accentuated adult mortality hump. This elevated mortality in the adult years is higher than in the other Asian families but remains well below that of any of the African patterns.

4.4.3 Family-specific Parameters

Tables 4.3 through 4.8 in Appendix A contain the component score vectors, median coefficients, and the below- and above-UFP deviations (constants) for males and females that when inserted into equation 4.1 produce the model life table schedules.

4.4.4 Evaluating System Performance

In order to evaluate the performance of this model, I perform two validation exercises. I first perform an in-sample validation similar to the in-sample validation described in the previous chapter. Using the model described in this chapter, I fit each of the 329 life tables in the INDEPTH dataset by first employing Discriminant Analysis to find the family that most closely approximates ${}_5q_0$ (probability of death before reaching age 5) and ${}_{45}q_{15}$ (the probability of death between age 15 and 60). I then select the level by finding the α value that creates a schedule that matches either life expectancy at birth (e_0), ${}_5q_0$, or ${}_{45}q_{15}$. I summarize the fit with the mean absolute error across age groups (0-75) and among all life table fits as well as the mean absolute error in e_0, ${}_5q_0$, and ${}_{45}q_{15}$.

Because this system is designed for use in developing regions, especially Africa, I compare the accuracy of estimation for the model described in this chapter with the UN model life tables for developing countries (United Nations, 1982). Although not designed specifically for developing settings other model life table systems like the Coale and Demeny and WHO modified logit model are still used to estimate mortality in developing settings, thus I include these for comparison as well. Likewise, the model recently developed by Wilmoth et al. (Wilmoth et al., 2012) produced remarkably accurate estimates when fitting the HMD data, so it too is included here for comparison. I use each of the four models to fit the in-sample 329 life tables and calculate the same summary measures.[7]

Results for INDEPTH model validation are presented in table 4.1. For all of the validation metrics, smaller numbers indicate better fit and the lowest value in each column is bolded. The system presented in this chapter does better on nearly every metric except for the mean absolute error for life expectancy for which the WHO and Wilmoth models perform slightly better when the level for the INDEPTH calibrated model is determined by child mortality or adult mortality. This seemingly paradoxical result of a smaller mean absolute error for life expectancy at birth but a larger mean absolute error across

[7]The UN system is fit using the procedure described on page 2 of Chapter IV of United Nations (1982), *Model Life Tables for Developing Countries* where I use the complete set of age-specific probabilities of death to estimate the appropriate loading factor. To fit the Coale and Demeny life tables I select the level by matching life expectancy to the closest half year and select the region that minimizes the sum of squared errors from the observed mortality rates. The WHO system was fit using the STATA software with the function modmatch, available at: `http://www.who.int/healthinfo/global_burden_disease/tools_software/en/`.The Wilmoth et al. model was fit with code provided at `http://www.demog.berkeley.edu/~jrw/LogQuad/`.

ages results because life expectancy is a function of the total area under the survivorship curve, which is determined by a set of age-specific mortality rates. As long as the area under the survivorship curve remains the same, any number of shapes can produce the same life expectancy at birth. In other words, any number of age specific mortality rate schedules can produce the same life expectancy at birth. Thus, it is possible to match life expectancy closely while missing the age-specific rates in any portion of the age-range if the area under the survival curve (l_x column of the life table) remains similar.

Figure 4.7, which shows the predicted fits from each of the four existing models and the three input combinations for the INDEPTH calibrated model for Agincourt females in 2007, demonstrates this paradoxical finding. When fitting a mortality curve with a large hump as is characteristic of high HIV prevalence populations like Agincourt, other models like the Wilmoth and WHO methods produce a high, flat set of mortality rates starting around age 20 through age 60. This pattern produces a similar life expectancy to the observed but misses the age-specific rates because the model essentially "averages" through the hump capturing the level of mortality but not the shape. The 'level' of mortality in the Wilmoth model is determined by the first input parameter, ${}_5q_0$, while the shape is determined by the second input parameter, 'k,' which adds a set of 'deviations' derived from an SVD of the residuals from a log-quadratic model that does not include 'k' and is optimized to match ${}_{45}q_{15}$ exactly. For Wilmoth, this high, flat shape is likely a function of the data used to calibrate his model as the HMD data do not contain mortality schedules with the high, concentrated HIV hump that results from very high prevalence. Nevertheless, because the Wilmoth model reproduces the input parameters exactly and, thus, closely matches the level of mortality over a large portion of the age range, it is able to accurately estimate an average mortality indicator like life expectancy at birth.

Despite the smaller MAE for life expectancy for the WHO and Wilmoth methods, the model presented in this chapter performs better amongst all age groups and life tables as shown by the smaller mean absolute error in the 'All-ages' column. When using life expectancy to determine the level, the model put forth in this chapter shows smaller errors compared to existing systems for both childhood and adult mortality measures.

Table 4.1: Mean absolute error for three mortality indicators and amongst all-ages and life tables after fitting INDEPTH life tables with various all-age mortality models.[†]

	Male				Female			
Model	All-ages[‡]	e_0	${}_5q_0$	${}_{45}q_{15}$	All-ages[‡]	e_0	${}_5q_0$	${}_{45}q_{15}$
INDEPTH MLTs								
level: e_0	**6.32**	—	**18.13**	**34.00**	**5.51**	—	**13.54**	**29.75**
level: ${}_5q_0$	7.42	3.20	—	76.78	6.01	2.50	—	56.29
level: ${}_{45}q_{15}$	6.69	2.62	32.34	—	6.11	2.42	24.28	—
UN	7.55	1.48	21.28	44.98	6.35	1.63	23.80	47.28
CD[§]	8.46	—	34.95	83.58	6.83	—	26.22	66.16
WHO[§]	7.84	**1.00**	—	—	6.78	**1.15**	—	—
Wilmoth[§]	8.03	1.35	—	—	6.73	1.68	—	—

[†] The mean absolute errors for 'All-ages', ${}_5q_0$, and ${}_{45}q_{15}$ are expressed per 1,000

[‡] 'All-ages' refers to the mean absolute error for the non-logged mortality rates across age groups (0, 1-4, 5-9, 10-14, ..., 75) and amongst all life tables $(1, ..., L = 312)$.

[§] 'WHO,' 'CD,' and 'Wilmoth' contain blank spaces as these quantities are inputs to these systems and thus have no error.

Naturally, any model calibrated with a certain dataset is likely to fit that dataset well compared to models that were calibrated with out-of-sample data. For this reason, I also complete a cross validation where I calibrate the model with a random 75% sample of the INDEPTH life tables and then fit the remaining 25%. I also fit the remaining 25% sample with the four existing model life table systems. This fitting procedure is performed iteratively 1,000 times yielding a distribution of 1,000 tables similar to tables 3.2 and 4.1. I report the mean errors from these distributions in table 4.2 below. Again, smaller numbers are better and the smallest number in each column is bolded.

Table 4.2: Mean of the distribution of mean absolute error for three mortality indicators and amongst all-ages and life tables after fitting INDEPTH life tables with various all-age mortality models after 1,000 iterations of cross validation.†

Model	Male				Female			
	All-ages‡	e_0	${}_5q_0$	${}_{45}q_{15}$	All-ages‡	e_0	${}_5q_0$	${}_{45}q_{15}$
INDEPTH MLTs								
level: e_0	**6.56**	—	**19.55**	**38.22**	**5.97**	—	**17.07**	**38.26**
level: ${}_5q_0$	7.71	3.41	—	83.78	6.51	3.11	—	71.47
level: ${}_{45}q_{15}$	7.06	2.86	35.26	—	6.60	2.95	30.94	—
UN	7.57	1.48	21.36	45.29	6.35	1.62	23.78	47.34
CD§	8.48	—	35.04	84.10	6.83	—	26.25	66.42
WHO§	7.84	**1.00**	—	—	6.77	**1.15**	—	—
Wilmoth§	8.03	1.36	—	—	6.73	1.69	—	—

† The mean absolute errors for 'All-ages', ${}_5q_0$, and ${}_{45}q_{15}$ are expressed per 1,000

‡ 'All-ages' refers to the mean absolute error for the non-logged mortality rates across age groups (0, 1-4, 5-9, 10-14, ..., 75) and amongst all life tables $(1, ..., L = 312)$.

§ 'WHO,' 'CD,' and 'Wilmoth' contain blank spaces as these quantities are inputs to these systems and thus have no error.

As with the cross validation for the HMD calibrated model, these results confirm the in-sample validation. The model life tables presented in this chapter largely outperform the other four models. Again, the model presented in this chapter, no matter how the level is determined, shows smaller errors across age compared to the three existing models. For males with the INDEPTH calibrated model, we see smaller errors in the 'All-ages' column when the level is determined by just childhood mortality or just adult mortality as well. For females with the INDEPTH model, errors are similar to those from the other four models when using child or adult mortality alone.

4.5 Discussion

In this chapter, I present a parsimonious mortality model 'calibrated' with data from the INDEPTH life table collection. The clustering method is able to identify empirical regularities in this collection of life tables with little input from the analyst and allows one to neatly summarize the characteristics of all-age mortality for the INDEPTH sites

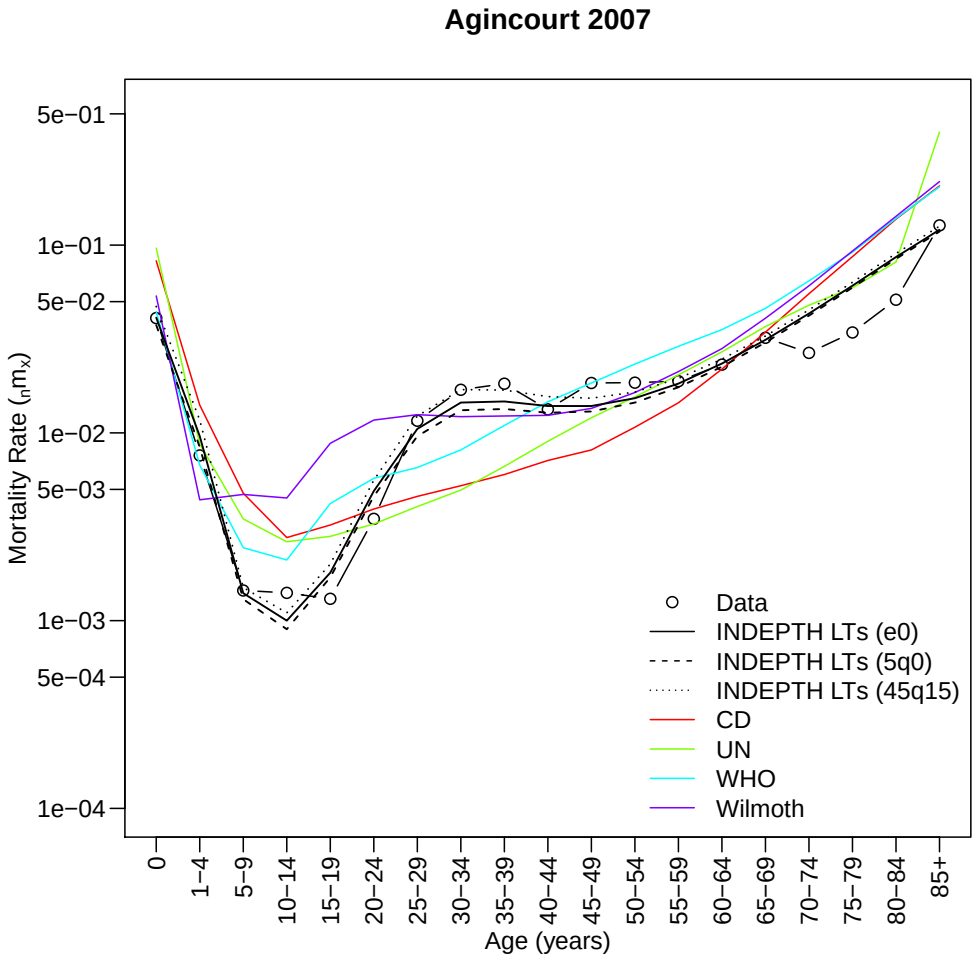


Figure 4.7: Fits of INDEPTH model with three different input combinations to determine the level for Agincourt HDSS Females 2007. 1.) Life expectancy at birth [solid black line] 2.) Child mortality [dashed black line] 3.) Adult mortality [dotted black line]. For comparison fits from the WHO modified logit model [teal solid line], Coale and Demeney model life tables [red solid line], UN model life tables for developing countries [green solid line], and the Wilmoth et al. model [purple solid line] are also shown.

included in this analysis.

To validate this system, I compare various measures of goodness of fit for our model and several existing age-specific mortality models. The INDEPTH calibrated model outperforms these other models on virtually all measures for both sexes. The INDEPTH patterns include ones with levels of adult mortality that do not appear in any existing model life table system – e.g. patterns 1 and 3, which represent mortality patterns generated under high HIV prevalence.

Perhaps most important, this system is easy to use and relatively flexible.[8] As with the HMD system, the user can approach the model life table system with a variety of mortality indicators with which one can identify the most similar family and appropriate level within that family. The discriminant analysis functionality is flexible and allows the user to train the system using an arbitrary combination of mortality indicators.

[8]To make this model life table system accessible and useful, I have created a package with a user-friendly graphical interface that implements the INDEPTH-calibrated version of this model for the free, open source statistical package `R` (R Development Core Team, 2011). The package can be requested from the author. It will likely also be posted to the Comprehensive R Archive Network for download in the future.

Appendix A - Fixed model parameters

Table 4.3: Component score vector values from INDEPTH calibration (left-singular vectors from SVD of INDEPTH data): **S**. When inserted into equation 4.1 with **B** from table 4.4 produces the underlying family patterns from the INDEPTH calibration.

	Male					Female				
Age	v1	v2	v3	v4	v5	v1	v2	v3	v4	v5
0	-0.10481	-0.24995	0.24287	-0.02180	0.34270	-0.10998	-0.28105	0.31225	-0.04536	0.32981
1-4	-0.17204	-0.28377	0.23728	0.36125	-0.15714	-0.17525	-0.30314	0.28551	0.40593	-0.05166
5-9	-0.21618	-0.18071	-0.12927	0.19733	-0.36149	-0.21887	-0.23914	-0.19926	0.16771	0.03269
10-14	-0.22563	-0.10375	-0.26341	0.07293	-0.29364	-0.22847	-0.19319	-0.24752	0.03852	0.22292
15-19	-0.22018	-0.00099	-0.31439	-0.00748	-0.16153	-0.21816	-0.07478	-0.21215	-0.05550	0.25018
20-24	-0.20909	0.11889	-0.23956	0.00543	-0.02850	-0.20524	0.06807	-0.05881	-0.06439	0.24004
25-29	-0.19698	0.23107	-0.09798	0.07428	0.05170	-0.19571	0.15556	0.09548	-0.04910	0.17639
30-34	-0.18641	0.28977	0.01708	0.15432	0.07160	-0.19028	0.17771	0.17957	-0.05771	0.07650
35-39	-0.17811	0.28164	0.07606	0.18607	0.06956	-0.18619	0.15500	0.18145	-0.08450	-0.01091
40-44	-0.17025	0.23824	0.07936	0.16920	0.06017	-0.18105	0.11123	0.16691	-0.11993	-0.10410
45-49	-0.16096	0.18119	0.06089	0.12405	0.04794	-0.17339	0.06206	0.16200	-0.16194	-0.18776
50-54	-0.14995	0.12610	0.03994	0.06565	0.03977	-0.16278	0.01185	0.16840	-0.20947	-0.22771
55-59	-0.13749	0.07943	0.01902	0.00939	0.03666	-0.14992	-0.03239	0.16845	-0.24762	-0.22921
60-64	-0.12410	0.04029	-0.00382	-0.04002	0.03489	-0.13553	-0.06942	0.14978	-0.26758	-0.19670
65-69	-0.11034	0.00677	-0.02886	-0.08275	0.03292	-0.12019	-0.09929	0.10679	-0.26392	-0.13161
70-74	-0.09671	-0.02284	-0.05459	-0.11626	0.03090	-0.10452	-0.12132	0.04122	-0.23649	-0.04141
75-79	-0.08371	-0.04803	-0.07825	-0.13826	0.03012	-0.08932	-0.13441	-0.03638	-0.19322	0.05687
80-84	-0.07183	-0.06576	-0.09640	-0.14716	0.03188	-0.07547	-0.13709	-0.10696	-0.14379	0.14147
85+	-0.06170	-0.07439	-0.10636	-0.14648	0.03348	-0.06392	-0.13236	-0.14940	-0.10285	0.18794

Table 4.4: Median coefficients for first five score vectors for INDEPTH calibration, by cluster/family, **B**. When inserted into equation 4.1 with **S** from table 4.3 produces the underlying family patterns for the INDEPTH calibration.

Cluster	Intercept	Coeff 1	Coeff 2	Coeff 3	Coeff 4	Coeff 5
1	0.4713	28.1187	3.2473	1.4817	1.3230	0.1349
2	0.0560	28.5974	-1.4503	0.2892	1.3416	-0.1916
3	-0.1364	27.7398	0.5984	0.9071	0.7073	0.0799
4	-0.5434	28.3542	-1.1758	0.4889	-1.7116	-0.4405
5	0.0626	31.8345	-1.3291	0.2911	-0.2650	0.2418
6	-0.5101	29.6758	-1.5352	0.0681	-1.8591	1.2890
7	0.5916	37.6681	2.2269	-2.4125	0.9689	-0.2876

Table 4.5: Male above-median family-specific deviations and cluster-invariant deviations for INDEPTH calibration: $\mathbf{D}_{f_+}$

	Family							
Age	1	2	3	4	5	6	7	$\mathbf{D}_{h_+}$
0	0.699	0.655	1.222	0.431	0.947	0.886	1.365	1.018
1-4	0.765	1.020	1.266	1.146	1.568	1.479	1.428	1.547
5-9	0.995	1.302	1.411	1.410	1.501	1.616	1.506	1.676
10-14	1.084	1.361	1.461	1.363	1.379	1.583	1.592	1.686
15-19	1.036	1.317	1.409	1.291	1.244	1.519	1.592	1.658
20-24	0.854	1.208	1.266	1.264	1.213	1.495	1.589	1.677
25-29	0.604	1.079	1.084	1.297	1.265	1.506	1.581	1.725
30-34	0.407	0.979	0.941	1.356	1.325	1.529	1.521	1.750
35-39	0.332	0.923	0.892	1.357	1.342	1.514	1.472	1.722
40-44	0.339	0.892	0.902	1.314	1.300	1.463	1.413	1.644
45-49	0.370	0.862	0.931	1.248	1.225	1.384	1.344	1.532
50-54	0.396	0.822	0.958	1.169	1.133	1.290	1.271	1.404
55-59	0.405	0.773	0.970	1.097	1.033	1.194	1.178	1.271
60-64	0.397	0.717	0.967	1.037	0.929	1.102	1.061	1.136
65-69	0.379	0.659	0.954	0.989	0.823	1.016	0.925	1.002
70-74	0.353	0.601	0.933	0.953	0.721	0.940	0.776	0.873
75-79	0.321	0.545	0.906	0.929	0.628	0.875	0.621	0.756
80-84	0.282	0.493	0.872	0.918	0.549	0.825	0.469	0.657
85+	0.235	0.445	0.832	0.919	0.489	0.791	0.336	0.580

Table 4.6: Male below-median family-specific deviations and cluster-invariant deviations for INDEPTH calibration: $\mathbf{D}_{f\_}$

	Family							
Age	1	2	3	4	5	6	7	$\mathbf{D}_{h\_}$
0	0.828	1.051	1.264	1.341	1.064	0.602	0.821	1.973
1-4	1.084	1.385	1.528	1.267	1.163	1.415	1.337	2.962
5-9	0.453	0.772	0.870	0.488	0.723	1.136	0.907	1.820
10-14	0.266	0.496	0.554	0.228	0.551	0.853	0.603	1.178
15-19	0.293	0.334	0.321	0.077	0.393	0.547	0.366	0.707
20-24	0.614	0.373	0.278	0.097	0.331	0.339	0.299	0.655
25-29	1.045	0.520	0.361	0.194	0.328	0.227	0.366	0.885
30-34	1.359	0.669	0.470	0.286	0.339	0.222	0.486	1.170
35-39	1.450	0.752	0.564	0.368	0.376	0.260	0.575	1.355
40-44	1.361	0.752	0.599	0.399	0.394	0.283	0.606	1.382
45-49	1.185	0.706	0.607	0.404	0.399	0.283	0.610	1.328
50-54	0.990	0.641	0.600	0.396	0.392	0.256	0.599	1.238
55-59	0.808	0.573	0.582	0.377	0.369	0.214	0.584	1.131
60-64	0.642	0.504	0.551	0.345	0.327	0.165	0.568	1.014
65-69	0.489	0.435	0.511	0.302	0.272	0.115	0.552	0.891
70-74	0.354	0.373	0.469	0.256	0.209	0.071	0.539	0.775
75-79	0.244	0.323	0.429	0.213	0.144	0.038	0.534	0.675
80-84	0.173	0.291	0.395	0.177	0.081	0.014	0.535	0.600
85+	0.137	0.273	0.369	0.147	0.025	0.000	0.543	0.553

Table 4.7: Female above-median family-specific deviations and cluster-invariant deviations for INDEPTH calibration: $\mathbf{D}_{f_+}$

	Family							
Age	1	2	3	4	5	6	7	$\mathbf{D}_{h_+}$
0	0.746	0.670	1.283	0.361	1.024	0.867	1.569	1.058
1-4	0.828	1.039	1.306	1.054	1.609	1.497	1.522	1.607
5-9	1.316	1.376	1.598	1.019	1.339	1.540	1.629	1.679
10-14	1.476	1.429	1.711	0.795	1.250	1.464	1.911	1.722
15-19	1.301	1.322	1.598	0.781	1.183	1.379	1.965	1.682
20-24	0.987	1.163	1.395	0.846	1.237	1.345	2.036	1.704
25-29	0.728	1.038	1.237	0.927	1.333	1.334	2.104	1.730
30-34	0.590	0.972	1.166	0.992	1.396	1.316	2.142	1.719
35-39	0.559	0.949	1.159	1.025	1.392	1.286	2.105	1.662
40-44	0.554	0.932	1.170	1.046	1.367	1.244	2.045	1.580
45-49	0.544	0.900	1.179	1.044	1.332	1.183	1.983	1.481
50-54	0.537	0.852	1.189	0.995	1.277	1.100	1.929	1.368
55-59	0.527	0.795	1.188	0.930	1.197	1.014	1.839	1.247
60-64	0.516	0.736	1.173	0.867	1.092	0.938	1.688	1.122
65-69	0.506	0.681	1.145	0.814	0.962	0.878	1.465	0.996
70-74	0.496	0.632	1.103	0.777	0.815	0.838	1.177	0.873
75-79	0.481	0.587	1.049	0.760	0.663	0.816	0.856	0.760
80-84	0.450	0.544	0.985	0.766	0.530	0.806	0.545	0.663
85+	0.397	0.500	0.919	0.790	0.439	0.801	0.302	0.588

Table 4.8: Female below-median Family-Specific deviations and cluster-invariant deviations for INDEPTH calibration: $\mathbf{D}_{f\_}$

	Family							
Age	1	2	3	4	5	6	7	$\mathbf{D}_{h\_}$
0	0.838	1.118	1.412	1.483	1.225	0.648	0.882	2.190
1-4	1.268	1.538	1.651	1.480	1.306	1.514	1.362	3.161
5-9	0.560	0.867	0.847	0.810	0.899	1.155	0.641	1.568
10-14	0.523	0.739	0.706	0.825	0.937	0.919	0.356	1.154
15-19	0.561	0.588	0.569	0.667	0.806	0.576	0.234	0.881
20-24	0.857	0.608	0.603	0.632	0.752	0.320	0.283	1.007
25-29	1.094	0.676	0.718	0.649	0.761	0.199	0.413	1.284
30-34	1.120	0.688	0.802	0.634	0.774	0.156	0.524	1.470
35-39	0.986	0.637	0.812	0.574	0.752	0.151	0.572	1.485
40-44	0.791	0.572	0.818	0.510	0.729	0.163	0.619	1.475
45-49	0.594	0.512	0.837	0.467	0.718	0.168	0.669	1.477
50-54	0.428	0.471	0.870	0.469	0.729	0.155	0.704	1.478
55-59	0.301	0.440	0.885	0.483	0.725	0.135	0.718	1.447
60-64	0.210	0.417	0.867	0.490	0.689	0.116	0.708	1.366
65-69	0.155	0.400	0.809	0.484	0.615	0.109	0.675	1.232
70-74	0.137	0.391	0.713	0.461	0.504	0.114	0.622	1.052
75-79	0.144	0.384	0.595	0.422	0.368	0.124	0.561	0.849
80-84	0.173	0.380	0.479	0.372	0.227	0.133	0.509	0.666
85+	0.205	0.379	0.398	0.323	0.113	0.133	0.489	0.553

Table 4.9: Alpha values to index family-specific 'levels' by life expectancy at birth for INDEPTH calibration, males

	Family						
Target e_0	1	2	3	4	5	6	7
30	0.826	0.980	1.086	1.631	1.392	1.618	1.705
32.5	0.712	0.906	1.005	1.551	1.314	1.534	1.609
35	0.592	0.832	0.923	1.473	1.241	1.455	1.516
37.5	0.467	0.754	0.841	1.395	1.161	1.372	1.425
40	0.335	0.673	0.758	1.316	1.082	1.290	1.337
42.5	0.180	0.590	0.672	1.232	1.000	1.204	1.247
45	-0.017	0.502	0.584	1.146	0.914	1.115	1.156
47.5	-0.141	0.409	0.495	1.059	0.828	1.024	1.066
50	-0.258	0.311	0.399	0.968	0.739	0.929	0.972
52.5	-0.374	0.201	0.301	0.873	0.641	0.831	0.876
55	-0.489	0.080	0.195	0.770	0.542	0.729	0.776
57.5	-0.610	-0.061	0.080	0.660	0.432	0.614	0.670
60	-0.735	-0.200	-0.063	0.539	0.316	0.499	0.557
62.5	-0.866	-0.334	-0.230	0.408	0.183	0.370	0.439
65	-1.004	-0.471	-0.389	0.264	0.041	0.233	0.307
67.5	-1.157	-0.614	-0.552	0.098	-0.188	0.080	0.160
70	-1.326	-0.776	-0.732	-0.142	-0.411	-0.205	-0.012
72.5	-1.513	-0.951	-0.920	-0.414	-0.640	-0.490	-0.329
75	-1.734	-1.161	-1.140	-0.691	-0.887	-0.769	-0.625
77.5	-1.994	-1.417	-1.397	-0.991	-1.178	-1.064	-0.944
80	-2.332	-1.750	-1.723	-1.366	-1.531	-1.424	-1.323
82.5	-2.798	-2.219	-2.181	-1.865	-2.030	-1.916	-1.836
85	-3.612	-3.044	-2.968	-2.716	-2.881	-2.754	-2.707

Table 4.10: Alpha values to index family-specific 'levels' by life expectancy at birth for INDEPTH calibration, females

	Family						
Target e_0	1	2	3	4	5	6	7
30	1.022	1.052	1.135	1.630	1.379	1.497	2.014
32.5	0.939	0.985	1.059	1.546	1.306	1.421	1.929
35	0.859	0.914	0.986	1.461	1.230	1.344	1.849
37.5	0.776	0.845	0.912	1.378	1.155	1.269	1.771
40	0.688	0.774	0.837	1.293	1.078	1.190	1.691
42.5	0.597	0.697	0.762	1.205	1.000	1.110	1.614
45	0.501	0.617	0.683	1.111	0.919	1.028	1.536
47.5	0.404	0.531	0.605	1.015	0.837	0.944	1.458
50	0.292	0.445	0.521	0.916	0.748	0.854	1.376
52.5	0.164	0.349	0.436	0.813	0.659	0.759	1.287
55	0.008	0.246	0.345	0.699	0.562	0.658	1.201
57.5	-0.143	0.124	0.249	0.573	0.457	0.550	1.111
60	-0.281	-0.014	0.149	0.438	0.349	0.433	1.017
62.5	-0.411	-0.150	0.040	0.285	0.229	0.307	0.913
65	-0.541	-0.283	-0.096	0.107	0.097	0.166	0.803
67.5	-0.682	-0.416	-0.239	-0.107	-0.068	0.000	0.686
70	-0.828	-0.558	-0.390	-0.321	-0.258	-0.308	0.556
72.5	-0.980	-0.713	-0.546	-0.521	-0.445	-0.545	0.410
75	-1.159	-0.887	-0.728	-0.746	-0.651	-0.781	0.245
77.5	-1.365	-1.098	-0.934	-0.985	-0.888	-1.029	0.051
80	-1.623	-1.365	-1.198	-1.280	-1.183	-1.334	-0.364
82.5	-1.973	-1.736	-1.563	-1.670	-1.589	-1.736	-0.853
85	-2.573	-2.388	-2.195	-2.320	-2.284	-2.408	-1.583

Chapter 5
CONCLUSION

This dissertation considers the use of models of age-specific mortality to characterize the age-specific risk of death during the course of the HIV epidemic in rural South Africa and to estimate complete sets of mortality rates based on model patterns when data on age-specific mortality are unavailable or limited. In this final chapter, I will synopsize the conclusions of the preceding chapters as well as summarize the implications and limitations of the work contained in this dissertation.

5.1 A 'Law of Mortality': The Heligman-Pollard Model

5.1.1 Summary

Chapter two was an application of the Heligman-Pollard model of age-specific mortality to data from a rural South African population covering the course of the HIV epidemic to characterize the likely sex-age-specific mortality effect of endemic HIV. The HP model fits well to the Agincourt data and the trend in the hump parameters reflects the likely impact of HIV. The level and spread parameters (D, E) show a steadily growing and gradually widening adult mortality hump. The hump parameters also reveal sex differences in the progression of mortality during the epidemic. For men, the epidemic started and remains 'older,' while for women mortality began concentrated in a narrow, young age range and expanded to older ages later. The results from this study of the Agincourt HDSS exhibit the profound impact of HIV on this population's mortality profile.

Chapter two not only contributes to the evidence linking HIV prevalence and age-specific mortality, in particular describing the intensity, shape, and location of the adult mortality hump resulting from likely HIV-related deaths, but also demonstrates the application of a robust estimation technique for a necessarily complex parametric model. The model was estimated using a Bayesian Melding method to produce probability distributions of the parameter values, model output ($_nq_x$), and corresponding life tables in four

periods during which the HIV epidemic grew rapidly in this population. From these distributions, one can make inference and confirm sizable and statistically significant declines in life expectancy during a rather brief time span. Previous work documented fit difficulties with the HP due to the high number of parameters, but the work in this dissertation shows it is possible to fit this kind of model given an appropriate fitting strategy.

5.1.2 Limitations

Although the model fits well to this data and confirms our understanding of the age-specific mortality impact of endemic HIV, this work contains a number of limitations. Based on the results of chapter two, I draw conclusions about the shape of mortality during an HIV epidemic, but this analysis is limited to just a single population with specific characteristics. First, the population under study here is located in a rural part of South Africa and may differ markedly from urban populations in the same country or rural areas elsewhere. Representativeness is also limited by the fact that the data for this study originate from a Health and Demographic Surveillance Site (Collinson, 2010). HDSS are closely monitored for health and demographic changes and are sometimes the testing ground for various interventions, which may influence health and mortality itself. Features of the epidemic itself including the timing of onset or the speed with which prevalence grew may also differ substantially from other settings. To the extent that these or some other features of the HDSS are distinct, these results may not generalize to other populations. Despite this limitation, age-specific mortality data are rare in Africa, so data of this type must be used to their full potential whenever possible to understand the causes of mortality patterns in Africa.

Another limitation of the analysis presented in chapter two is that cause of death is not known for the deaths in this study. The Bayesian estimation method produces probabilistic estimates of uncertainty around the various mortality indicators that suggest these results are unlikely to have occurred by chance, but without cause of death data, there is no direct link between age-specific mortality rates and HIV prevalence.

5.1.3 Future Research

As noted above, cause of death data is necessary to establish a direct link between HIV and elevated adult mortality in Africa. Although not available at the date of this dissertation, cause of death data for Agincourt should be available in the future. Work in this area should prioritize methods for ascertaining cause of death in settings like Agincourt including use of verbal autopsy, a method for ascertaining cause of death where an interviewer uses a questionnaire to document the symptoms and demographics of a deceased person from a household member or someone familiar with the deceased (Snow et al., 1992; Chandramohan et al., 1994; Kahn et al., 2000; Baiden et al., 2007) with some mechanism for automated assignment of cause of death (Byass et al., 2003, 2006; Fantahun et al., 2006; Murray et al., 2007), which is typically done by consensus of two or more physicians. If cause of death is established, one would expect HIV-related causes to dominate and could link those causes to change in the HP parameters.

One of the limitations mentioned above is that of representativeness in that the data come from an HDSS. Agincourt is just one of nearly four dozen such HDSS organized under the umbrella of the "International Network for the Demographic Evaluation of Populations and Their Health in Developing Countries" (INDEPTH, 2011; Bangha et al., 2010). Generalizability issues may be as least partially resolved by adding analysis from other HDSS, and if the HP could be fit to other sites experiencing endemic HIV, it may be possible to identify a typical set of changes in the Heligman-Pollard parameters that is strongly related to increases in HIV-related deaths. Furthermore, one could model these typical sets of parameters as a function of any number of variables that could alter the course of the HIV epidemic such as the availability of highly active anti-retroviral therapies, which lengthen life for infected individuals.

Bayesian Melding as an estimation technique for highly parameterized models may also be applied to other demographic models. Not only does the Bayesian approach used here avoid many of the pitfalls of estimating highly parameterized models, it also provides a means for assessing uncertainty. For instance, one could investigate fertility change in a similar fashion by fitting a parametric model of age-specific fertility within a Bayesian framework.

5.2 Model Life Tables

5.2.1 Summary

Chapters three and four present a new method for model life table construction. Using the statistical technique model-based clustering, I identify similar age patterns of mortality among two collections of life tables covering both high and low-to-middle income regions. The clustering method is almost entirely automated and requires very little input from the analyst allowing the patterns to reveal themselves. The clustering results also reveal a story of mortality and epidemiological change in these two collections. From the Human Mortality Database, which covers 37 mostly developed countries, five typical age patterns emerge, which roughly correspond to stages of the epidemiological transition. While time was the primary dimension along which the HMD life tables clustered, the comparatively shorter data series containing arguably more rapid mortality change among the INDEPTH tables cluster along geographic and epidemiological conditions. Treating the clusters as 'families' in a traditional model life table system, a one-parameter model is proposed to adjust the 'level' of mortality within each family. The result is a new effectively two-parameter model life table system for the countries and time periods contained in these two collections. For each calibration of the model life table system, HMD and INDEPTH, an in-sample validation and cross validation were performed and errors for several mortality indicators were compared to those obtained from several existing model life table systems. The HMD-calibrated model performs similarly to the Wilmoth model, which was also calibrated with HMD data (although a slightly smaller collection of tables, 719), and the WHO modified logit model, while the INDEPTH-calibrated model showed smaller errors on nearly every metric for virtually all input combinations.

Given that such a substantial proportion of countries around the globe still rely on model life table systems to estimate mortality in the absence of even partial vital registration systems (Gerland, 2009; Mathers et al., 2005), updating and improving the demographic methods we use to quantify the risk of death should be a priority for this field. A primary contribution from chapter three is just that. The model presented in that chapter based on data from the Human Mortality Database includes patterns not available in the widely-used, existing systems because those conditions did not exist at

the times those systems were created. Likewise, the model described in chapter four calibrated with data from INDEPTH includes numerous life tables from various African populations, something no other model life table system can claim. From these data, patterns emerge that are not seen in any existing model including an 'HIV pattern' to depict the accentuated adult mortality hump. Taken together, these two models can produce mortality rate schedules over a wide range of shapes and levels that no other system can.

5.2.2 Limitations

As with all model life table systems, this model is limited by the data that went into it. Although both datasets are high quality and cover an array of mortality levels and shapes, it would be misguided to expect these models to accurately reflect mortality resulting from conditions not present in the countries and time periods contained in these two datasets. Likewise, despite the INDEPTH system having a pattern which likely reflects the effect of HIV, there is no direct introduction of HIV prevalence as a parameter to alter the age-specific mortality profile.

The life table construction method presented in this dissertation follows a traditional mode life table system structure, but recent models have moved away from the "family/level" organization in favor of more continuous models (Wilmoth et al., 2012; Murray et al., 2003). Unlike the traditional setup where the user would first select a family then level, these models take one or two input parameters (usually child morality) and exploit the relationship between child mortality and mortality at other ages to produce a complete set of mortality rates. The recent model by Wilmoth et al. produces the most accurate estimates of all model life table systems to date and does so with a two continuous parameter configuration with the second parameter being optional. The first parameter determines the level and the second, 'k,' then alters the resulting schedule to determine the shape. Using two continuous parameters removes the one discrete/one continuous arrangement of the Coale and Demeny and UN systems, which can be sensitive to the choice of a regional model (Preston et al., 2001) (the discrete parameter). Although Wilmoth calculates errors for life expectancy that are nearly twice as large when using just one input parameter, ${}_5q_0$) (Wilmoth et al., 2012), if one has the requisite mortality indicators, the Wilmoth method will produce estimates with small errors for high-income

country data under its current calibration. Essentially, when enough data is available the Wilmoth model performs very well. Although the system presented here is able to capture a wide range of mortality shapes, I do not actually extrapolate between patterns, which would approximate a more continuous model.

5.2.3 Future Research

Future research in this area should address the limitations listed above. I will first address the issue of combining patterns for a more continuous model. As stated in chapter three, the user of one of these systems can use Discriminant Analysis to first select the appropriate family. The `R` package, `mclust`, which was used to conduct the cluster analysis provides functions to perform DA where one can classify cases into the values of a categorical dependent variable based on the known classification of other observations. These functions also provide a likelihood of classification into each cluster. By weighting the fixed parameters that govern the output from these models, it is possible to estimate a complete set of age-specific mortality rates based on more than one pattern. In this case, one could cover nearly every possible mortality shape and level contained in these data. This improvement in the method would not only produce possibly better fits but it gains the attractive qualities of Wilmoth's continuous structure while maintaining the traditional 'family/level' configurations, which can be more intuitive and useful if one lacks information on one of the mortality indicators that serve as input parameters to the Wilmoth model. The family/level structure allows for the introduction of qualitative information or expert opinion in selecting the appropriate age pattern of mortality in the absence of knowledge of certain input parameters. For instance, one may not have a reliable estimate of ${}_5q_0$ or ${}_{45}q_{15}$, but may know something about epidemiological conditions or about neighboring populations, which could guide the selection of a family or pattern. Also, unlike the Wilmoth model, which cannot change the shape without a second input parameter, the method presented in chapters three and four could determine both the shape and level with just a single mortality indicator. Wilmoth's method also shows near zero error for the input parameters because the model matches the input quantities exactly, which the method presented in this dissertation does not. Peforming an optimization of either the deviations added to underlying family pattern or to the weights

for each SVD component such that the model output reproduces the input quantities exactly could also prove helpful in further reducing error for the method presented in the preceding chapters.

Another avenue for future work is the generation of a model of age-specific mortality specifically for countries with generalized HIV epidemics. For the purpose of population projection, mortality conditions play a key role in future population composition thus the mortality component of projections for countries with generalized epidemics should include the influence of the future trajectory of the epidemic. A model based on data from countries with generalized epidemics that incorporates HIV prevalence (and perhaps anti-retroviral therapy coverage) as a parameter could go a long way in accurately projecting the mortality portion of a population forecast. Using tools like the United Nations Estimation and Projections Package (Brown et al., 2010; Ghys and Garnett, 2010) one can estimate or project features of an HIV epidemic such as prevalence and coverage of anti-retroviral therapy among the infected population, which could then be converted to a set of future age-specific mortality rates with a model specifically for high-HIV prevalence countries. This can be easily accomplished by modeling the coefficients that result from regressing the observed mortality schedules on the SVD components. The coefficients concentrate the age-variation in mortality into a small number of coefficients, which can be modeled by any number of socially, economically, or demographically relevant variables. In the case of the HMD data, for instance, which clusters primarily along a temporal dimension thus mirroring the stages of the epidemiological transition, one could determine the most predictive variables of mortality at the various stages of transition.

5.3 Conclusion

Accurately and parsimoniously modeling age-specific mortality remains an important and effective tool for demographers. This dissertation demonstrates that, when applied to longitudinal data from a population experiencing high HIV prevalence and estimated with a robust, innovative method, the Heligman-Pollard model can explicate the effect of a rapidly expanding HIV epidemic on the age profile of mortality. Also, given the need for simple, reliable means of estimating mortality rates in the absence of high quality vital registration systems, model life table systems persist as a crucial tool for demographers,

other mortality analysts, and policy makers alike. Barring the sudden implementation of full vital registrations systems, upgrading models like these are the best way to understand a population's mortality landscape. Better understanding of the age pattern of mortality should help guide present and future policy affecting population health and distribution of societal resources.

BIBLIOGRAPHY

Adjuik, M., T. Smith, S. Clark, J. Todd, A. Garrib, Y. Kinfu, K. Kahn, M. Mola, A. Ashraf, H. Masanja, K. Adazu, J. Sacarlal, N. Alam, A. Marra, A. Gbangou, E. Mwageni and F. Binka. 2006. "Cause-specifc mortality rates in sub-Saharan Africa and Bangladesh." *Bulletin of the World Health Organization* 84(3):181–188.

Alkema, L., A. E. Raftery and S. J. Clark. 2007. "Probabilistic Projections of HIV Prevalence Using Bayesian Melding." *The Annals of Applied Statistics* 1(1):229–248.

Andreev, E. M., E. Nolte, V. M. Shkolnikov, E. Varavikova and M. McKee. 2003. "The evolving pattern of avoidable mortality in Russia." *International Journal of Epidemiology* 32(3):437–446.

Baiden, F., A. Bawah, S. Biai, F. Binka, T. Boerma, P. Byass, D. Chandramohan, S. Chatterji, C. Engmann, D. Greet et al. 2007. "Setting international standards for verbal autopsy." *Bulletin of the World Health Organization* 85:570–571.

Banfield, J. and A. Raftery. 1993. "Model-based Gaussian and non-Gaussian clustering." *Biometrics* 49(3):803–821.

Bangha, M., A. Diagne, A. Bawah and O. Sankoh. 2010. "Monitoring the millennium development goals: the potential role of the INDEPTH Network." *Global health action* 3.

Bebbington, M., C.-D. Lai and R. Zitikis. 2007. "Modeling human mortality using mixtures of bathtub shaped failure distributions." *Journal of Theoretical Biology* 245:528–538.

Black, R., S. Cousens, H. Johnson, J. Lawn, I. Rudan, D. Bassani, P. Jha, H. Campbell, C. Walker, R. Cibulskis, T. Eisele, L. Liu, C. Mathers and Child Health Epidemiology Reference Group of WHO and UNICEF. 2010. "Global, regional, and national causes of child mortality in 2008: a systematic analysis." *Lancet* 375(9730):1969–87.

Blacker, J. 2004. "The impact of AIDS on adult mortality: evidence from national and regional statistics." *AIDS* 18(suppl. 2):S19–S26.

Bobak, M. and M. Marmot. 1996. "East-West mortality divide and its potential explanations: Proposed research agenda." *BMJ: British Medical Journal (International Edition)* 312(7028).

Botha, J. and D. Bradshaw. 1985. "African Vital Statistics - A Black Hole?" *South African Medical Journal* 67:977–981.

Brass, W. 1971. "On the Scale of Mortality." In *Biological Aspects of Demography*, edited by W. Brass, London: Taylor and Francis.

Brown, T., L. Bao, A. Raftery, J. Salomon, R. Baggaley, J. Stover and P. Gerland. 2010. "Modeling HIV epidemics in the antiretroviral era: the UNAIDS Estimation and Projection package 2009." *Sexually transmitted infections* 86:3–10.

Byass, P., E. Fottrell et al. 2006. "Refining a probabilistic model for interpreting verbal autopsy data." *Scandinavian journal of public health* 34(1):26.

Byass, P., D. Huong and H. Van Minh. 2003. "A probabilistic approach to interpreting verbal autopsies: methodology and preliminary validation in Vietnam." *Scandinavian Journal of Public Health* 31:32–37.

Carriere, J. 1992. "Parametric models for life tables." *Insurance: Mathematics and Economics Insurance: Mathematics and Economics* 14(1).

Chandramohan, D., G. Maude, L. Rodrigues and R. Hayes. 1994. "Verbal autopsies for adult deaths: issues in their development and validation." *International journal of epidemiology* 23(2):213.

Clark, S. J. 2002. "INDEPTH Mortality Patterns for Africa." In *Population and Health in Developing Countries*, vol. 1, edited by INDEPTH network, Ottawa: IDRC Press.

Clark, S. J., M. Jasseh, S. Punpuing, E. Zulu, O. S. Ayaga Bawah and INDEPTH Network Member Sites. 2009. "INDEPTH Model Life Tables 2.0." Presented at the Annual Meeting for the Population Association of America.

Clark, S. J., J. Thomas and L. Bao. 2010. "Estimates of Age-Specific Reductions in HIV Prevalence in Uganda: Bayesian Melding Estimation with an HIV-enabled Cohort Component Projection Model." *In Preparation* .

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

Coale, A. J. and G. Guo. 1989. "Revised regional model life tables at very low levels of mortality." *Population Index* 55(4):613–643.

Collinson, M. A. 2010. "Striving against adversity: the dynamics of migration, health and poverty in rural South Africa." *Global health action* 3.

Congdon, P. 1993. "Statistical Graduation in Local Demographic Analysis and Projection." *Journal of the Royal Statistical Society: Series A* 156(2):237–270.

Dabis, F. and E. R. Ekpini. 2002. "HIV-1/AIDS and maternal and child health in Africa." *Lancet* 359:2097–2104.

De Cock, K. M., M. G. Fowler, E. Mercier, I. de Vincenzi, J. Saba, E. Hoff, D. J. Alnwick, M. Rogers and N. Shaffer. 2000. "Prevention of mother to child HIV transmission in resource poor countries - translating research into policy and practice." *Journal of the American Medical Association* 283:1175–1182.

Debón, A., F. Montes and R. Sala. 2005. "A comparison of parametric models for mortality graduation. Application to mortality data for the Valencia Region (Spain)." *SORT* 29(2):269–288.

Dellaportas, P. 2001. "Bayesian Analysis of Mortality Data." *Journal of the Royal Statistical Society: Series A* 164(2):275–291.

Deparcieux, A. 1746. *Essai sur les Probabiliés de la Durée de la Via Humaine*. Paris: Guérin Frères.

Department of Health, Republic of South Africa. 1995. "Fifth national HIV survey in women attending antenatal clinics of the public health services in South Africa, Oct/Nov." *Epidemiological Comments* 22.

———. 1996. "Sixth national HIV survey of women attending antenatal clinics of the public health services in the Republic of South Africa, Oct/Nov 1995." *Epidemiological Comments* 23:3–15.

———. 1997. "Seventh national HIV survey of women attending antenatal clinics of the public health services, October/November 1996." *Epidemiological Comments* 23(2):4–16.

———. 1998. "Eighth annual national HIV sero-prevalence survey of women attending antenatal clinics in South Africa, 1997." .

———. 2003. "National HIV and Syphilis Antenatal Sero-Prevalence Survey in South Africa: 2002." URL `http://www.doh.gov.za/docs/reports/2002/hiv-syphilis.pdf`.

———. 2006. "National HIV and Syphilis Antenatal Sero-Prevalence Survey in South Africa: 2005." URL `http://www.doh.gov.za/docs/reports/2005/hiv.pdf`.

Doherty, T., D. McCoy and S. Donohue. 2005. "Health system constraints to optimal coverage of the prevention of mother-to-child HIV transmission programme in South Africa: lessons from the implementation of the national pilot programme." *African health sciences* 5(3):213.

Dorrington, R., D. Bourne, D. Bradshaw, R. Laubscher and I. M. Timæus. 2001. "The Impact of HIV/AIDS on Adult Mortality in South Africa." Tech. rep., Burden of Disease Research Unit, Medical Research Council. URL `http://www.mrc.ac.za/bod/`.

Fantahun, M., E. Fottrell, Y. Berhane, S. Wall, U. Högberg and P. Byass. 2006. "Assessing a new approach to verbal autopsy interpretation in a rural Ethiopian community: the InterVA model." *Bulletin of the World Health Organization* 84:204–210.

Forfar, D. and D. Smith. 1987. "The Changing shape of English Life Tables." *Transactions of the Faculty of Actuaries* 40:98–133.

Foster, G. and J. Williamson. 2000. "A review of current literature of the impact of HIV/AIDS on children in sub-Saharan Africa." *AIDS* 14(suppl. 3):S275–S284.

Fraley, C. and A. Raftery. 2002. "Model-based clustering, discriminant analysis, and density estimation." *Journal of the American Statistical Association* 97(458):611–631.

———. 2006. "Mclust version 3 for R: Normal mixture modeling and model-based clustering." Tech. Rep. 504, Department of Statistics, University of Washington.

Gage, T. and C. Mode. 1993. "Some laws of mortality: how well do they fit?" *Human biology; an international record of research* 65(3):445.

Gerland, P. 2009. "Estimation of robust demographic time-series with noisy data for developing countries: selected case studies from Africa and Asia." Presented at the Center for Statistics and the Social Sciences Seminar, University of Washington.

Ghys, P. D. and G. P. Garnett. 2010. "The 2009 HIV and AIDS estimates and projections: methods, tools and analyses." *Sexually transmitted infections.* 86(2):ii1.

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.

Graunt, J. 1662. *Natural and political observations made upon the bills of mortality*. London: Roycroft and Dicas.

Groenewald, P., N. Nannana, D. Bourne, R. Laubscher and D. Bradshaw. 2005. "Idetiying deaths from AIDS in South Africa." *AIDS* 19:193–201.

Guo, G. 1993. "Mortality Trends and Causes of Death: A Comparison between Eastern and Western Europe, 1960s-1980s." *European Journal of Population / Revue Européenne de Démographie* 9(3):287–312.

Halley, E. 1693. "An Estimate of the Degrees of the Mortality of Mankind." *Philosophical Transactions of the Royal Society of London* 17:596–610.

Hartmann, M. 1987. "Past and Recent Attempts to Model Mortality at All Ages." *Journal of Official Statistics* 3(1):19–36.

Heligman, L. and J. H. Pollard. 1980. "The Age Pattern of Mortality." *Journal of the Institute of Actuaries* 107(434):49–80.

Horiuchi, S. and A. J. Coale. 1982. "A Simple Equation for Estimating the Expectation of Life at Old Ages." *Population Studies* 36(2):317–326.

Hosegood, V., A.-M. Vaenneste and I. M. Timæus. 2004. "Levels and causes of adult mortality in rural South Africa: the impact of AIDS." *AIDS* 18:663–667.

INDEPTH. 2011. "INDEPTH – International Network for the Demographic Evaluation of Populations and Their Health in Developing Countries." URL `http://www.indepth-network.org`.

Jaffar, S., A. Grant, J. Whitworth, P. Smith and H. Whittle. 2004. "The natural history of HIV-1 and HIV-2 infections in adults in Africa: a literature review." *Bulletin of the World Health Organization* 82:462–469.

Kahn, K., S. Tollman, M. Collinson, S. Clark, R. Twine, B. Clark, M. Shabangu, F. Gomez-Olive, O. Mokoena and M. Garenne. 2007. "Research into health, population and social transitions in rural South Africa: Data and methods of the Agincourt Health and Demographic Surveillance System." *Scandinavian Journal of Public Health* 35(3):8.

Kahn, K., S. Tollman, M. Garenne and J. Gear. 2000. "Validation and application of verbal autopsies in a rural area of South Africa." *Tropical Medicine & International Health* 5(11):824–831.

Karim, Q. A. and S. S. A. Karim. 1999. "Epidemiology of HIV Infection in South Africa." *AIDS* 13:4–7.

Kostaki, A. 1992. "A nine-parameter version of the Heligman-Pollard formula." *Mathematical population studies* 3(4):277–88.

Lopez, A. D., C. D. Mathers, M. Ezzati, D. T. Jamison and C. J. L. Murray. 2006. "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data." *Lancet* 367(9524).

Lynch, S. M. and J. S. Brown. 2005. "A New Approach to Estimating Life Tables with Covariates and Constructing Interval Estimates of Life Table Qunatities." *Sociological Methodology* 35:189–237.

Makeham, W. H. 1860. "On the law of mortality and the construction of annuity tables." *Journal of the Institute of Actuaries* 8:301–310.

Marshall, S. J. 2004. "Developing countries face double burden of disease." *Bulletin of the World Health Organization* 82(7).

Marston, M., J. Todd, J. R. Glynn, K. E. Nelson, R. Rangsin, T. Lutalo, M. Urassa, S. Biraro, L. V. der Paal, P. Sonnenberg and B. Żaba. 2007. "Estimating 'net' HIV-related mortality and the importance of background mortality rates." *AIDS* 21(suppl 6):S65–S71.

Marston, M., B. Zaba, J. A. Salomon, H. Brahmbhatt and D. Bagenda. 2005. "Estimating the Net Effect of HIV on Child Mortality in African Populations Affected by Generalized HIV Epidemics." *Journal of Acquired Immune Deficiency Syndrome* 38:219–227.

Mathers, C. D., D. Ma Fat, M. Inoue, C. Rao and A. D. Lopez. 2005. "Counting the Dead and What They Died From: An Assessment of the Global Status of Cause of Death Data." *Bulletin of the World Health Organization* 83(3).

McKee, M. and V. Shkolnikov. 2001. "Understanding the toll of premature death among men in eastern Europe." *BMJ: British Medical Journal (International Edition)* 323(7320).

McKeown, T. 1976. *The modern rise of population*. New York: Academic Press.

———. 1979. *The role of medicine : dream, mirage, or nemesis?*. Princeton, N.J.: Princeton University Press.

McNown, R. and A. Rogers. 1989. "Forecasting Mortality: A Parameterized Time Series Approach." *Demography* 26(4):645–660.

Morgan, D., C. Mahe, B. Mayanja, J. Okongo, R. Lubega and J. Whitworth. 2002. "HIV-1 infection in rural Africa: is there a difference in median time to AIDS and survival compared with that in industrialized countries?" *AIDS* 16(4):597.

Murray, C., B. Ferguson, A. Lopez, M. Guillot, J. Salomon and O. Ahmad. 2003. "Modified Logit Life Table System: Principles, Empirical Validation, and Application." *Population Studies* 57:165–182.

Murray, C. and A. Lopez. 1996. "The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries and Risk Factors in 1990 and Projected to 2020." In *The Global Burden of Disease and Injury series*, vol. 1, Cambridge, MA: Harvard University Press.

Murray, C., A. Lopez, D. Feehan, S. Peter and G. Yang. 2007. "Validation of the symptom pattern method for analyzing verbal autopsy data." *PLoS Med* 4(11):e327.

Murray, C. J. L. and A. D. Lopez. 1997. "Mortality by cause for eight regions of the world: Global Burden of Disease Study." *Lancet* 349(9061).

Murray, C. J. L., L. C. Rosenfeld, S. S. Lim, K. G. Andrews, K. J. Foreman, D. Haring, N. Fullman, M. Naghavi, R. Lozano and A. D. Lopez. 2012. "Global malaria mortality between 1980 and 2010: a systematic analysis." *Lancet* 379(9814).

Newell, M.-L., H. Brahmbhatt and P. D. Ghys. 2004. "Child mortality and HIV infection in Africa: a review." *AIDS* 18(suppl. 2):S27–S34.

Notzon, F., Y. Komarov, S. Ermakov, C. Sempos, J. Marks and E. Sempos. 1998. "Causes of declining life expectancy in Russia." *JAMA : the journal of the American Medical Association* 279(10):793–800.

Nyirenda, M., V. Hosegood, T. Bärnighausen and M.-L. Newell. 2007. "Mortality levels and trends by HIV serostatus in rural South Africa." *AIDS* 21(suppl 6):S73–S79.

Olshansky, S. J. and A. B. Ault. 1986. "The Fourth Stage of the Epidemiologic Transition: The Age of Delayed Degenerative Diseases." *The Milbank Quarterly* 64(3):355–391.

Omran, A. 1971. "The epidemiologic transition: a theory of the epidemiology of population change." *The Milbank Memorial Fund Quarterly* 49:509–38.

Perks, W. 1932. "On Some Experiments In The Graduation Of Mortality Statistics." *Journal of the Institute of Actuaries* 63(1):12–57.

Poit, P., M. Bartos, P. D. Ghys, N. Walker and B. Schwartländer. 2001. "The global impact of HIV/AIDS." *Nature* 410:968–973.

Poole, D. and A. E. Raftery. 2000. "Inference for Deterministic Simulation Models: The Bayesian Melding Approach." *Journal of the American Statistical Association* 95(452):1244–1255.

Porter, K. and B. Zaba. 2004. "The empirical evidence for the impact of HIV on adult mortality in the developing world: data from serological studies." *AIDS* 18(suppl. 2):S9–S17.

Preston, S. H. 1976. *Mortality Patterns in National Populations with Special Reference to Recorded Causes of Death*. New York: Academic Press.

Preston, S. H., P. Heuveline and M. Guillot. 2001. *Demography: Measuring and Modeling Population Processes*. Malden, MA: Blackwell.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`, ISBN 3-900051-07-0.

Raftery, A. E. and L. Bao. 2010. "Estimating and Projecting Trends in HIV/AIDS Generalized Epidemics Using Incremental Mixture Importance Sampling." *Biometrics* 66:in press.

Roger, A. 1987. "Parameterized multistate population dynamics and projections." *Journal of the American Statisitcal Association* 81:48–61.

Rogers, A. 1986. "Parameterized multistate population dynamics and projections." *Journal of the American Statistical Association* 81(393):48–61.

Rogers, A. and K. Gard. 1991. "Applications of the Heligman/Pollard Model Mortality Schedule." *Population Bulletin of the United Nations* 30:79–105.

Rubin, D. 1988. "Using the SIR algorithm to simulate posterior distributions." In *Bayesian Statistics 3: proceedings of the third valencia international meeting, June 1-5, 1987*, edited by J. Bernardo, M. DeGroot, D. Lindley and A. Smith, Clarnedon Press.

Rubin, D. B. 1987. "A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputation When Fractions of Missing Information are Modest: The SIR Algorithm." *Journal of the American Statistical Association* 82(398):543–546.

Salomon, J. A. and C. J. L. Murray. 2002. "The Epidemiologic Transition Revisited: Compositional Models for Causes of Death by Age and Sex." *Population and Development Review* 28(2):205–228.

Sharrow, D. J. 2011. *HPbayes: Heligman Pollard mortality model parameter estimation using Bayesian Melding with Incremental Mixture Importance Sampling*. URL `http://CRAN.R-project.org/package=HPbayes`, r package version 0.1.

Shkolnikov, V. M. and G. A. Cornia. 1998. "Causes of the Russian Mortality Crisis: Evidence and Interpretations." *World Development* 26(11).

Siler, W. 1979. "A Competing-Risk Model for Animal Mortality." *Ecology* 60(4):750–757.

———. 1983. "Parameters of mortality in human populations with widely varying life spans." *Statistics in medicine* 2(3).

Smith, J., P. Mushati and F. Kurwa. 2007. "Changing patterns of adult mortality as the HIV epidemic matures in Manicaland, eastern Zimbabwe." *AIDS* 21(suppl 6):S81–S86.

Snow, R., M. Winstanley, V. Marsh, C. Newton, C. Waruiru, I. Mwangi, P. Winstanley, K. Marsh, D. Forster and J. Armstrong. 1992. "Childhood deaths in Africa: uses and limitations of verbal autopsies." *The Lancet* 340(8815):351–355.

Steele, R. J., A. E. Raftery and M. J. Edmonds. 2006. "Computing Normalizing Constants for Finite Mixture Models via Incremental Importance Sampling (IMIS)." *Journal of Computational and Graphical Statistics* 15(3):712–734.

Steele, R. J., A. E. Raftery and M. J. Emond. 2003. "Computing Normalizing Constants for Finite Mixture Models via Incremental Mixture Importance Sampling (IMIS)." Tech. Rep. 436, Department of Statistics, University of Washington, Seattle, Washington.

Timæus, I. M. and M. Jasseh. 2004. "Adult Mortality in Sub-Saharan Africa: Evidence from Demographic and Health Surveys." *Demography* 41(4):757–772.

Todd, J., J. R. Glynn, M. Marston, T. Lutalo, S. Biraro, W. Mwita, V. Suriyanon, R. Rangsin, K. E. Nelson, P. Sonnenberg, D. Fitzgerald, E. Karita and B. Żaba. 2007. "Time from HIV seroconversion to death: a collaborative analysis of eight studies in six low and middle-income countries before highly active antiretroviral therapy." *AIDS* 21(suppl 6):S55–S63.

Tollman, S. M., K. Kahn, M. Garenne and J. S. Gear. 1999. "Reversal in mortality trends: evidence from the Agincourt field site, South Africa, 1992– 1995." *AIDS* 13:1091–1097.

UNAIDS. 2012. "2012 Report on Global AIDS Epidemic."

UNAIDS and WHO. 2008. "2008 Report on the Global AIDS Epidemic." Tech. rep., UNAIDS, WHO.

———. 2009. "AIDS Epidemic Update: December 2009." Tech. rep., UNAIDS, WHO.

United Nations. 1982. "Model Life Tables for Developing Countries." In *Population Studies*, 77, New York: United Nations.

———. 2004. *World Population Prospects: the 2002 Revision*, vol. 3. New York: United Nations.

University of California, Berkeley and Max Planck Institute for Demographic Research. Data Downloaded November 2009. "Human Mortality Database." URL `http://www.mortality.orgorwww.humanmortality.de`.

Vaupel, J. W., K. G. Manton and E. Stallard. 1979. "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality." *Demography* 16(3):pp. 439–454. URL `http://www.jstor.org/stable/2061224`.

Ševčíková, H., A. E. Raftery and P. A. Waddell. 2007. "Assessing Uncertainty in Urban Simulations Using Bayesian Melding." *Transportation Research* 41B(6):652–669.

Wilmoth, J., K. Andreev, D. Jdanov and D. Glei. 2007. "Methods Protocol for the Human Mortality Database Version 5." www.mortality.org/Public/Docs/MethodsProtocol.pdf.

Wilmoth, J., V. Canudas-Romo, S. Zureick, M. Inoue and C. Sawyer. 2009. "A Flexible Two-Dimensional Mortality Model for Use in Indirect Estimation." Presented at the Annual Meeting for the Population Association of America.

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

Wood, J. W., D. J. Holman, K. A. O'Conner and R. J. Ferrell. 2001. "Mortality Models for Paleodemography." Tech. rep., in Center for Studies in Demography and Ecology Working Paper Series, University of Washington.

World Health Organization. 2011. "World Malaria Report 2011." Tech. rep., World Health Organization, Geneva.

Zaba, B., M. Marston, A. C. Crampin, R. Isingo, S. Biraro, T. Bärnighausen, B. Lopman, T. Lutalo, J. R. Glynn and J. Todd. 2007. "Age-specific mortality patterns in HIV-infected individuals: a comparative analysis of Africa community study data." *AIDS* 21(suppl 6):S87–S96.

Zaba, B., A. Whiteside and J. T. Boerma. 2004. "Demographic and socioeconomic impact of AIDS: taking stock of the empirical evidence." *AIDS* 18(suppl 2):S1–S7.