

# More on the Cohort-Component Model of Population Projection in the Context of HIV/AIDS: A Leslie Matrix Representation and New Estimation Methods

Jason Thomas & Samuel J. Clark<sup>1</sup>

Last Modified: 2008-11-20

<sup>1</sup>Department of Sociology, University of Washington (Jason Thomas and Samuel J. Clark PhD), Institute of Behavioral Science (IBS), University of Colorado at Boulder, CO, USA and MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, University of the Witwatersrand, Johannesburg, South Africa (Samuel J. Clark). Correspondence to Samuel J. Clark: [samclark@u.washington.edu](mailto:samclark@u.washington.edu). The project described here was supported by a seed grant from the Center for Statistics and the Social Sciences (CSSS), University of Washington and by grant numbers 1 K01 HD057246-01 and 1 R01 HD054511 01 A1 from the National Institute of Child Health and Human Development (NICHD) of the National Institutes of Health (NIH). The content of the work presented here is solely the responsibility of the authors and does not necessarily represent the official views of the the NIH. The authors are grateful to Patrick Heuveline for his correspondence and for making his work available. We also thank Leontine Alkema, Le Bao, Jennifer Chunn, Jeff Eaton, Marie-Louise Newell and Adrian Raftery for invaluable discussion during the preparation of this work. The first author would also like to thank the Africa Centre for Health and Population Studies, University of KwaZulu-Natal for their hospitality and useful comments on an early version of this work.

## **Abstract**

This article presents an extension of the cohort component model of population projection (CCMPP) first formulated by Heuveline (2003) that is capable of modeling a population affected by HIV. Heuveline proposes a maximum likelihood approach to estimate the age profile of HIV incidence that produced the HIV epidemics in East Africa during the 1990s. We extend this work by developing the Leslie matrix representation of the CCMPP which greatly facilitates the implementation of the model for parameter estimation and projecting. The Leslie matrix also contains information about the stable tendencies of the corresponding population, such as the stable age distribution and time to stability. We validate our reformulation of the model by comparing parameter estimates obtained through maximum likelihood and bootstrap methods to those presented by Heuveline using Heuveline's original data set. A further application of the model to a small population with high HIV prevalence in rural South Africa is presented as an additional demonstration. This work lays the foundation for development of more robust and flexible Bayesian estimation methods that will greatly enhance the utility of this and similar models.

KEY WORDS: Model, Cohort-Component, Leslie Matrix, Estimation, HIV, Incidence, Prevalence, Africa.

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>4</b>
<b>2</b>	<b>CCMPP</b>	<b>5</b>
2.1	The HIV Incidence Trend . . . . .	7
2.2	Additional HIV-related Force of Mortality . . . . .	9
2.3	Matrix Notation for HIV-enabled CCMPP . . . . .	13
<b>3</b>	<b>PARAMETER ESTIMATION</b>	<b>16</b>
3.1	Types of Data . . . . .	17
3.2	Likelihoods . . . . .	18
3.2.1	HIV Test Results in a General-Population Sample . . . . .	19
3.2.2	HIV Test Results in an ANC-Patient Sample . . . . .	19
3.2.3	HIV Test Results in all or a Sample of Births from HIV+ Mothers . . . . .	20
3.2.4	HIV Test Results During a Follow-up of an HIV- Sample . . . . .	20
3.2.5	Survival During a Follow-up of HIV+ Individuals . . . . .	21
3.3	Estimation Techniques . . . . .	22
<b>4</b>	<b>RESULTS</b>	<b>22</b>
<b>5</b>	<b>DISCUSSION</b>	<b>31</b>
	<b>REFERENCES</b>	<b>32</b>

## List of Tables

1	Five-Year Incidence Rates Calculated from the Gamma Density and an Exponential Curve. . . . .	8
2	Expected Number of Person-Years Lived Over a Five-Year Interval by Survival Schedule and Duration Group. . . . .	10
3	Survival Probabilities Applied to HIV+ ( $s_{a,d>1}$ ). . . . .	12
4	Life Table for HIV+ Women Infected at Age 15. . . . .	13
5	Estimates of CCMPP Parameters and 95% Confidence Intervals. . . . .	24
6	Comparison of Model Fit to Data on Age-Specific HIV Prevalence from Kapiri Mposhi, Zambia. . . . .	25
7	CCMPP Parameter Estimates and 95% Confidence Intervals using Data from KwaZulu-Natal, South Africa. . . . .	27

## List of Figures

1	Age-specific HIV incidence rates, $i_{a,t}$ , for different values of the population-specific scale parameter, $H$ , over time. . . . .	6
2	Projected Prevalence Over Time for Women Aged 15-49. . . . .	28
3	Projected Prevalence Over Time for Men Aged 15-54. . . . .	29
4	Projected Population Pyramids at 25 Year Intervals for KwaZulu-Natal South Africa. . . . .	30

# 1 INTRODUCTION

The cohort component model of population projection (CCMPP) is perhaps *the* iconic method in demography, see for example Bowley (1924); Cannan (1895); Whelpton (1936); Leslie (1945); Pritchett (1891); Pearl and Reed (1920); Dorn (1950). This classic method increments forward in time a population defined by age according to a specified life table and set of age-specific fertility rates, taking into account net migration at each age. In its basic form it is straightforward and easy to implement which has allowed it to become one of the essential tools used by governments and planning organizations to help them understand the likely future size and composition of a population, and how that may change under different assumptions or as a result of interventions of various types.

Fundamentally the CCMPP relates the age structure of a population to fertility, mortality and migration, with the current age structure being the result of fertility, mortality and migration at each age in the past. Most commonly a future age structure of the population is ‘predicted’ given a time series of age-specific fertility, mortality, and net migration. The model can also be used to *estimate* trends in a subset of the four components given the others, and it is a use of this type that occupies us here.

The HIV epidemic affecting Africa and other parts of the developing world poses significant challenges to demographers concerned with either measuring the current state of an affected population or predicting its future. Because HIV affects both fertility and mortality in important ways (for example: Ford and Hosegood, 2005; Garenne et al., 2007; Gregson et al., 2007; Hunter et al., 2003; Kahn et al., 2007; Lewis et al., 2004; Nyirenda et al., 2007; Terceira et al., 2003; Timaes and Jasseh, 2004; Wachter et al., 2002; Zaba and Gregson, 1998; Zaba et al., 2007; Carpenter et al., 1997; Gray et al., 1998; Nunn et al., 1997; Sewankambo et al., 1994; Todd et al., 1997), it is not possible to understand the dynamics of a population affected by HIV without specifically taking into account these effects. Further complicating this situation, data describing underlying fertility and mortality unaffected by HIV are scarce and often of poor quality, especially for most populations with high HIV prevalence.

These challenges are addressed in an interesting and useful way in a model developed by Heuveline (2003). Heuveline created a multi-state version of the standard CCMPP model (Day, 1996; United Nations, 2004) that further classifies the population by time since infection with HIV and uses a set of age-specific incidence parameters to ‘infect’ HIV negative people and transition them to the first (shortest duration) HIV positive group. This version of the model also includes additional parameters to govern the links between HIV status and fertility. Heuveline used data describing HIV status and survival (mortality) from East African countries to estimate these model parameters using maximum likelihood techniques.

This article provides a detailed description of Heuveline’s multi-state, HIV-enabled CCMPP and two different procedures used to estimate its parameters, the maximum likelihood method employed by Heuveline and an approach based on bootstrap methods. The first section provides a detailed introduction to the CCMPP and presents the Leslie matrix representation that we have created for it. Following is a description of the data and methods used to estimate the model parameters. Next, to validate both our new implementation and Heuveline’s original work, we present our new parameter estimates obtained from an analysis

of the same data used by Heuveline in his original work. Finally, we apply the model to a rural population in South Africa with high HIV prevalence and present estimates of the key parameters describing that population.

## 2 CCMPP

Heuveline (2003) extends the standard CCMPP to accommodate a population categorized by duration of infection with HIV using five ‘HIV duration’ groups. There are four HIV+ duration groups (0-4 years, 5-9 years, 10-14 years, and 15+ years) as well as an HIV- group. In this section we present Heuveline’s multi-state CCMPP for a population with 17 age groups (0-4, 5-9, . . . , 80+) in each of the five HIV duration groups. The model is introduced with a series of equations representing the transition from one group/time period to the next. While the model can be applied to both men and women, the description presented here only includes the details for women.

Begin by dividing the population into age groups where  $a = 1, 2, \dots, 17$  correspond to age groups 0-4, 5-9, . . . , 80+. Denote membership in the HIV duration groups by  $d$ , with  $d = 1, 2, \dots, 5$  corresponding to HIV-, HIV+ for 0-4 years, . . . , HIV+ for more than 15 years. Time is indexed by  $t$  noting that the duration between  $t$  and  $t + 1$  is equal to the width of a standard age interval, e.g. 5 years. Let  $n_{a,d,t}$  be the number of women in age group  $a$  and duration group  $d$  at time  $t$ . For  $1 < a < 17$ , we have:

$$n_{a+1,1,t+1} = n_{a,1,t} s_{a,1,t} (1 - i_{a,t}) \quad (1)$$

$$n_{a+1,2,t+1} = n_{a,1,t} s_{a,1,t} i_{a,t} s_{a,2} \quad (2)$$

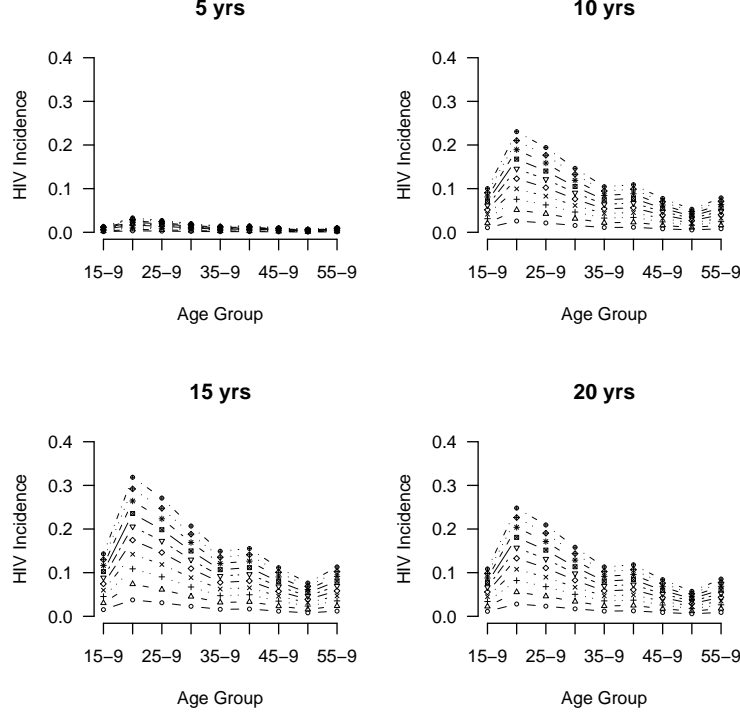
$$n_{a+1,d,t+1} = n_{a,d-1,t} s_{a,1,t} s_{a,d} \quad \text{for } d > 2 \quad (3)$$

where  $s_{a,d}$  is the survivorship ratio for age group  $a$  and duration group  $d$ . Note that for  $2 < d < 5$  this survivorship ratio determines the transition from one age group to the next, as well as from one duration group to the next. Each HIV+ group is exposed to the same underlying base survivorship ratio as the HIV- group in addition to this extra survivorship ratio that accounts for the increased mortality associated with different durations of infection. The parameter  $i_{a,t}$  is the fraction of women in age group  $a$  who become infected with HIV over the projection interval. To allow for the heterogeneity of HIV epidemics across populations, this parameter is decomposed as:

$$i_{a,t} = 1 - \exp \{ -\Gamma_{t-t_0} H j_a \} \quad (4)$$

where  $\Gamma_{t-t_0}$  is a parametric curve used to model the time trend in the HIV epidemic from the start time  $t_0$ . The actual values for  $\Gamma_{t-t_0}$  are presented in Table 1 (see the next section for more details). The parameter  $H$  is a population-specific scale parameter that captures the overall magnitude of the epidemic. The parameter  $j_a$  is an age- and sex-specific scaling factor for incidence that represents the multiplicative difference in HIV incidence between age group  $a$  and a reference age group which is held constant at a value of 1.0 in order to make the model identifiable. Following Heuveline we set the reference age group to 25-29, i.e.  $j_5 = 1.0$ .

Figure 1: Age-specific HIV incidence rates,  $i_{a,t}$ , for different values of the population-specific scale parameter,  $H$ , over time.



For a given age profile of incidence (a specific set of values for  $j_a$ ), Figure 1 demonstrates how the different values for  $H$  simple scale the incidence profile. Each panel in this figure corresponds to a different time in the epidemic with incidence whose overall scale is determined by the values of  $\Gamma$ . Within each panel, each line corresponds to a value of the population-specific scale parameter  $H$  ranging from 0.1 to 1.0.

The projection equations are slightly different for the youngest and oldest age groups. The oldest (open-ended) age group is incremented by two sources, those 75-79 and 80+ in the previous time period. Thus for  $a = 17$  we have:

$$n_{17,1,t+1} = n_{16,1,t} s_{16,1,t} (1 - i_{16,t}) + n_{17,1,t} s_{17,1,t} (1 - i_{17,t}) \quad (5)$$

$$n_{17,2,t+1} = n_{16,1,t} s_{16,1,t} i_{16,t} s_{16,2} + n_{17,1,t} s_{17,1,t} i_{17,t} s_{17,2} \quad (6)$$

$$n_{17,d,t+1} = n_{16,d-1,t} s_{16,1,t} s_{16,d} + n_{17,d-1,t} s_{17,1,t} s_{17,d} \quad \text{for } 2 < d < 5 \quad (7)$$

$$n_{17,5,t+1} = n_{16,4,t} s_{16,1,t} s_{16,5} + n_{17,4,t} s_{17,1,t} s_{17,5} + n_{16,5,t} s_{16,1,t} s_{16,5} + n_{17,5,t} s_{17,1,t} s_{17,5} \quad (8)$$

As with the single-state CCMPP, the number of children in the first age group at the end of the projection interval is determined by surviving forward the births that occur during the projection interval. The number of births that occur is calculated by applying age-specific fertility rates to the average number of women in each age group during the projection interval, taking into account the fact that HIV+ women who have been infected for different durations will to varying degrees be less likely to have children. To capture the relationship between fertility and HIV status, Heuveline defined three additional parameters. First, consider the number of HIV- births:

$$n_{1,1,t+1} = s_{0,1,t} \frac{1}{1 + SRB} \times \left( \sum_{a=\alpha}^{\beta} f_{a,1,t} \frac{n_{a,1,t} + p_{a-1,1,t}^- n_{a-1,1,t}}{2} + \sum_{d=2}^5 \sum_{a=\alpha}^{\beta} f_{a,d,t}^- \frac{n_{a,d,t} + p_{a-1,d-1,t} n_{a-1,d-1,t}}{2} \right) \quad (9)$$

In Equation 9 above, the  $f_{a,1,t}$ 's are simply the age-specific fertility rates for HIV- women, and the lower and upper bounds of the childbearing age range are  $\alpha$  and  $\beta$ . Fertility among HIV+ women introduces the following parameters

$$f_{a,d,t}^- = f_{a,1,t} e_a g_d (1 - v) \quad (10)$$

for  $1 < d$ . The superscript in  $f_{a,d,t}^-$  designates HIV- births to women who are HIV+ (i.e.  $d > 1$ ). The parameter  $v_d$  is the probability that an HIV+ woman in duration group  $d$  will give birth to an HIV+ child, the *vertical transmission* rate. The parameter  $e_a$  captures the higher level of sexual activity and resulting fertility among HIV+ women age 15-19 who have been infected for 0-4 years ( $d = 2$ ), accounting for the selection of more sexually active teenagers into the HIV+ category. In other words we expect  $e_{a=4} > 1$  while  $e_{a \neq 4}$  are constrained to be 1.0. The parameter  $g_d$  represents the *fertility impairment* experienced by women in duration group  $d$ , a number that becomes smaller as the time since infection increases reflecting increasing fertility impairment with time since infection. The corresponding equations for HIV+ births are:

$$n_{1,2,t+1} = s_{0,1,t} \frac{1}{1 + SRB} \sum_{d=2}^5 \sum_{a=\alpha}^{\beta} f_{a,d,t}^+ \frac{n_{a,d,t} + p_{a-1,d-1,t} n_{a-1,d-1,t}}{2} \quad (11)$$

$$f_{a,d,t}^+ = f_{a,1,t} e_a g_d v \quad (12)$$

Finally, we define the factors used to approximate the average number of women at the beginning and end of the period,  $p_{a,1,t}^-$  and  $p_{a,d,t}$ :

$$p_{a,1,t}^- = s_{a,1,t} (1 - i_{a,t}) \quad (13)$$

$$p_{a,1,t} = s_{a,1,t} i_{a,t} s_{a,2} \quad (14)$$

$$p_{a,d,t} = s_{a,1,t} s_{a,d} \quad \text{for } d > 1. \quad (15)$$

## 2.1 The HIV Incidence Trend

Age-specific incidence in CCMPP is modeled as follows:

$$i_{a,k,t} = 1 - \exp \{ -\Gamma_{t,t_0} H j_{a,k} \} \quad (16)$$

Table 1: Five-Year Incidence Rates Calculated from the Gamma Density and an Exponential Curve.

<b>Time Period</b>	$\Gamma_{t,t_0}$	$E_{t,t_0}$
0 - 5 years	0.028	0.063
6 - 10 years	0.216	0.191
11 - 15 years	0.316	0.323
16 - 20 years	0.235	0.457
20+ years <sup>a</sup>	0.163	0.540

<sup>a</sup>For the gamma model it is assumed that the HIV incidence rate will level off at the rate equal to the integral of the gamma density from 20 and 21 multiplied by five. For the exponential model it is assumed that the HIV incidence rate will level off at a rate equal to  $5 * (h(t = 21) - h(t = 20))$ . See the text for the definition of  $h(t)$ .

where  $i_{a,k,t}$  is the fraction of individuals age  $a$  who will become infected over the projection interval.  $\Gamma_{t,t_0}$  represents the shape of the incidence trend from the start of the epidemic in year  $t_0$  to the projection period  $t$ . This incidence trend is shifted up or down by  $H$ , an overall scale parameter for the epidemic. Finally,  $j_{a,k}$  is the incidence rate ratio comparing those of age  $a$  and sex  $k$  to women age 25-29. The details of the incidence trend  $\Gamma_{t,t_0}$  are described in this section, along with several other possible specifications.

The incidence trend  $\Gamma_{t,t_0}$  used by Heuveline (2003) is borrowed from EpiModel, a computer program developed by the World Health Organization to make short-term projections of adult AIDS cases (Chin and Lawnga, 1991) and the precursor of UNAIDS’ estimation and projection package (EPP) software currently used to estimate the prevalence of HIV (Ghys et al., 2004).  $\Gamma_{t,t_0}$  is based on the gamma family of distributions:

$$g(t) = \frac{t^{\alpha-1} e^{-t/\beta}}{(\alpha-1)! \beta^\alpha}, \quad \text{for } t \geq 0, \alpha > 0, \beta > 0 \quad (17)$$

The  $\alpha$  parameter is typically referred to as the shape parameter since it affects how peaked or flat the density is, as  $\alpha$  increases the density appears more flat or uniform. The scale parameter  $\beta$  is associated with how diffuse or spread out the density is, as  $\beta$  increases the density spreads out.

Before discussing the calculations made by Heuveline (2003) it is helpful to discuss a particular property of the gamma distribution. The mode, or value for which the the function reaches its maximum, is equal to  $(\alpha - 1)\beta$ . This quantity has a nice interpretation in that it is the number of years after the start of the epidemic  $t_0$  when the epidemic peaks. For example, if the shape parameters is 5 and the scale parameter is 3, then the epidemic peaks 12 years after it began<sup>1</sup>.

This is precisely the density used by Heuveline (2003) to calculate the five-year incidence rates, i.e.  $\alpha = 5, \beta = 4$ . The actual rates are calculated by integrating the gamma density

---

<sup>1</sup>Chin and Lawnga (1991) report that setting  $\alpha = 5$  results in “the best empirical ‘fit’ to the reported AIDS-case curves in countries with reliable case-reporting systems.” They set  $\beta = 1$ .

over the appropriate five-year span, i.e:

$$\Gamma_{t-t_0} = \int_{t-5}^t \frac{t^{5-1} e^{-x/4}}{(5-1)!4^5} dt \quad \text{for } t \in \{5, 10, 15, 20\} \quad (18)$$

$$\Gamma_{t-t_0} = 5 \times \int_{20}^{21} \frac{t^{5-1} e^{-x/4}}{(5-1)!4^5} dt \quad \text{for } t > 20 \quad (19)$$

The five-year incidence rate for twenty years after the start date is different because the decline in the gamma density for values greater than 20 is too rapid to represent an actual decline in incidence. The values used to estimate the CCMPP parameters are presented in Table 1. We are aware that this is relatively crude, and it would be preferable to model this trend using time series techniques with attention given to the uncertainty around the trend, especially at time points far into the future. That is among the improvements that currently occupy us.

We will now discuss two other possible specifications for the trend in HIV incidence. The first is an exponential curve that models a continual increase in HIV over time. While this may not be realistic in the long run it does provide an upper bound for the trend. A reasonable lower bound is a constant rate of new infections (i.e. no increase) over time. Since the second specification is simply a constant<sup>2</sup> we will focus our attention on the exponential model.

The exponential curve used to model the trend in HIV incidence takes the following functional form:

$$h(t) = \frac{e^{\beta t}}{\beta} - t \quad \text{for } t = 1, 2, 3, \dots; \quad \text{and } \beta > 0 \quad (20)$$

The five-year HIV incidence rates are calculated by differencing  $h(t)$  (at lag one) and summing over the five year period of interest:

$$E_{t,t_0} = \sum_{j=1}^5 h(t) - h(t-1) \quad (21)$$

Heuveline (2003) chooses a value of  $\beta = 0.005$  to obtain the five-year HIV incidence rates based on the exponential model, shown in Table 1. Finally, similar to the gamma model described above, the incidence rate after twenty years takes a different form. For the exponential curve,  $E_{20+,t_0} = 5 \times (h(t=21) - h(t=20))$ .

## 2.2 Additional HIV-related Force of Mortality

In the HIV-enabled CCMPP individuals infected with HIV ( $d > 1$ ) experience an additional force of mortality that is not experienced by those in the HIV- state ( $d = 1$ ). This mortality differential can be seen in the following projection equations

$$n_{a+1,d=2,t+1} = n_{a,d=1,t} s_{a,d=1,t} i_{a,t} s_{a,d=2} \quad (22)$$

$$n_{a+1,d>2,t+1} = n_{a,d-1,t} s_{a,d=1,t} s_{a,d>2} \quad (23)$$

---

<sup>2</sup>Heuveline (2003) uses 0.2 as the five-year incidence rate.

Table 2: Expected Number of Person-Years Lived Over a Five-Year Interval by Survival Schedule and Duration Group.

Duration Group	Survival Schedule			
	3	8	11	12
0 to 0-4 (d=2)	2.7750	4.7100	4.8000	4.8310
0-4 to 5-9 (d=3)	0.4250	2.4300	3.3750	3.6000
5-9 to 10-4 (d=4)	0.0000	0.8600	2.0000	2.4125
10-4 to 15-9 (d=5)	0.0000	0.3150	1.0000	1.5375

where  $s_{a,d>1} < 1$ ; survival in the HIV+ states is reduced compared to the HIV- state. Recall that the vital rates (mortality and fertility) in CCMPP are treated as fixed parameters and need to be set by the user.

Heuveline turns to the epidemiological literature for guidance on choosing values for the survival rates of individuals infected with HIV. One of the more important findings (concerning the HIV-enabled CCMPP) is that the progression from HIV to death is faster for older individuals. Morgan et al. (2002) report this finding in a study of a cohort from rural Uganda for whom the time of HIV infection is reasonably well known. Their data include 10, 18, and 19 deaths among 65, 68, and 35 participants in age groups 15-24, 25-39, and 40+ years. The cumulative probability of survival for each group is 79% (95%CI: 63-88%), 72% (95%CI: 56-83%), and 20% (95%CI: 6-40%), respectively. Morgan et al. (2002) also report a faster progression from seroconversion to AIDS for the oldest age group (40+ years).<sup>3</sup>

Given the evidence from the epidemiological literature, Heuveline (2003) specifies the survival rates for HIV+ individuals as a function of age at infection. This dependence comes through in the choice of a particular survival schedule, defined by the median number of years lived after infection. The schedules include median survival times of 3, 8, 10, and 12 years. Children who are infected perinatally follow the 3 year schedule, while the oldest age groups follow the 8-year schedule. Before describing the age dependence in greater detail, it is helpful to take a slight digression and define some more notation.

Heuveline (2003) defines these survival schedules with reference to the projection interval (i.e. five years). Let  $y_{d,m}$  be the expected number of person-years lived by an individual in duration group  $d$  following survival schedule  $m$ , where  $m = 3, 8, 11, 12$ . For example, the average number of years lived by a person infected 5-9 years ago who is following the survival schedule with a median survival time of 11 years is  $y_{d=3,m=11} = 3.375$ . The values for  $y_{d,m}$  (see Table 2) are estimated from distributions of median survival time of those infected with HIV (Heuveline, 2003; Chin, 1994).

Now we are in a position to define  $s_{a,d>1}$ . Let us begin with those who have been infected for 0-4 years ( $d = 2$ ). Children born and infected (perinatally) during the projection interval are exposed to:

$$s_{a=1,d=2} = \frac{y_{d=2,m=3}}{5} \quad (24)$$

<sup>3</sup>In this same study, the median time from seroconversion to AIDS is 9.4 years, IQR: 5.5 – 10.1 years. The median time from seroconversion to death is 9.8 years, IQR: 6.1 – > 10.3 years (Morgan et al., 2002).

The HIV-related survival ratios for the next two age groups are defined to be 1.0 because persons between the of ages 5 and 9 are not able to be infected given our current assumptions about incidence, recall that the age-specific incidence rates are zero for the first three age groups. (This might not be the best assumption since infants are still able to become infected via breastfeeding.) For those who are ages 15-19, 25-34, or above age 45, the additional force of mortality caused by HIV takes a form similar to the equation just above:

$$s_{a=4,d=2} = \frac{y_{d=2,m=12}}{5} = 0.9662 \quad (25)$$

$$s_{8 \geq a \geq 6,d=2} = \frac{y_{d=2,m=11}}{5} = 0.9600 \quad (26)$$

$$s_{a \geq 10,d=2} = \frac{y_{d=2,m=8}}{5} = 0.9420 \quad (27)$$

Note how survival declines as age increases before leveling off at age 45 and older. For the age groups not already mentioned, the survival parameters are calculated by taking the average over two adjacent survival schedules:

$$s_{a=5,d=2} = \frac{y_{d=2,m=11} + y_{d=2,m=12}}{2} = 0.9631 \quad (28)$$

$$s_{a=9,d=2} = \frac{y_{d=2,m=8} + y_{d=2,m=11}}{2} = 0.9651 \quad (29)$$

We now turn our attention to the third duration group, individuals who have been infected for 5-9 years. For this group, we start with those aged 5-9:

$$s_{a=2,d=3} = \frac{y_{d=3,m=3}}{y_{d=2,m=3}}. \quad (30)$$

The corresponding parameters for the older age groups take a similar form. The expected number of person-years lived for the third duration group is divided by the expected number of person-years lived by the second duration group (for a given survival schedule). The dependence on age for  $d = 3$  takes the same form as for the previous duration group, only that the age groups are incremented by one. The actual equations are:

$$s_{a=5,d=3} = \frac{y_{d=3,m=3}}{y_{d=2,m=3}} \quad (31)$$

$$s_{a=6,d=3} = \frac{y_{d=3,m=12} + y_{d=3,m=11}}{y_{d=2,m=12} + y_{d=2,m=11}} \quad (32)$$

$$s_{9 \geq a \geq 7,d=3} = \frac{y_{d=3,m=11}}{y_{d=2,m=11}} \quad (33)$$

$$s_{a=10,d=3} = \frac{y_{d=3,m=11} + y_{d=3,m=8}}{y_{d=2,m=11} + y_{d=2,m=8}} \quad (34)$$

$$s_{a \geq 11,d=3} = \frac{y_{d=3,m=8}}{y_{d=2,m=8}} \quad (35)$$

The pattern continues for the fourth and fifth duration groups. All of the parameters for the additional force of mortality due to HIV are listed in Table 3.

Life tables can also be constructed using the survival rates presented in Table 3, where cohorts defined by age at infection are exposed to the survival rates aligned along the diagonal cells

Table 3: Survival Probabilities Applied to HIV+ ( $s_{a,d>1}$ ).

Age Group	HIV Duration Group			
	0-4 yrs ( $d = 2$ )	5-9 yrs ( $d = 3$ )	10-4 yrs ( $d = 4$ )	15+ yrs ( $d = 5$ )
0-4	0.5550	–	–	–
5-9	–	0.1532	–	–
10-4	–	–	0.0000	–
15-9	0.9662	–	–	0.0000
20-4	0.9631	0.7452	–	–
25-9	0.9600	0.7242	0.6701	–
30-4	0.9600	0.7031	0.6326	0.6373
35-9	0.9600	0.7031	0.5926	0.5751
40-4	0.9510	0.7031	0.5926	0.5000
45-9	0.9420	0.6104	0.5926	0.5000
50-4	0.9420	0.5159	0.4927	0.5000
55-9	0.9420	0.5159	0.3539	0.4598
60-4	0.9420	0.5159	0.3539	0.3663
65-9	0.9420	0.5159	0.3539	0.3663
70-5	0.9420	0.5159	0.3539	0.3663
75-9	0.9420	0.5159	0.3539	0.3663
80+	0.9420	0.5159	0.3539	0.3663

Table 4: Life Table for HIV+ Women Infected at Age 15.

Age Group	${}_n p_x$	${}_n q_x$	$l_x$	${}_n d_x$	${}_n L_x$	$T_x$	${}^0 e$
15	0.9566	0.0434	100,000	4,345	489,138	1,612,020	16.1202
20	0.7347	0.2653	95,655	25,381	414,822	1,122,882	11.7389
25	0.6587	0.3413	70,274	23,983	291,412	708,060	10.0757
30	0.6255	0.3745	46,291	17,338	188,110	416,648	9.0006
35	0.5634	0.4366	28,953	12,642	113,160	228,538	7.8934
40	0.4883	0.5117	16,311	8,346	60,690	115,378	7.0736
45	0.4863	0.5137	7,965	4,092	29,595	546,88	6.8660
50	0.4823	0.5177	3,873	2,005	14,352	25,093	6.4789
55	0.4370	0.5630	1,868	1,052	6,710	10,741	5.7498
60	0.3388	0.6612	816	540	2,730	4,031	4.9395
65	0.3220	0.6780	276	187	912	1,301	4.7125
70	0.2940	0.7060	89	63	288	389	4.3669
75	0.2500	0.7500	26	20	80	101	3.8713
80	0.1586	0.8414	6	6	21	21	3.4425

of the table. For example children infected by their mothers will never reach age 15 years because the survival probability is zero for those infected at birth and between the ages of 10 and 14. The mortality experienced by the cohort of women infected at age 15 is summarized in the life table presented in Table 4.

The size of the cohort exposed to the mortality risks in Table 4 is reduced to half after fifteen years. After thirty years there is just under ten percent of the cohort still living. While the survival experience of this cohort may seem plausible, it is preferable to have their survival rates informed by data. The current specification for the survival experience results in a stable population with unappealing characteristics. This point is discussed later on in the paper.

## 2.3 Matrix Notation for HIV-enabled CCMPP

These equations for the multi-state, HIV-enabled CCMPP can be conveniently expressed in matrix notation. For a population with 17 age groups and five HIV duration groups, the

population at time  $t$  is represented by an  $85 \times 1$  column vector:

$$\mathbf{n}_t = \begin{bmatrix} n_{1,1,t} \\ n_{2,1,t} \\ \vdots \\ n_{17,1,t} \\ \hline \vdots \\ n_{1,4,t} \\ n_{2,4,t} \\ \vdots \\ n_{17,4,t} \end{bmatrix} \quad (36)$$

The corresponding *Leslie* matrix is:

$$\mathbf{A}_t = \begin{bmatrix} \mathbf{B}_{1,1} & \mathbf{B}_{1,2} & \mathbf{B}_{1,3} & \mathbf{B}_{1,4} & \mathbf{B}_{1,5} \\ \mathbf{B}_{2,1} & \mathbf{B}_{2,2} & \mathbf{B}_{2,3} & \mathbf{B}_{2,4} & \mathbf{B}_{2,5} \\ \mathbf{0} & \mathbf{B}_{3,2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_{4,3} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_{5,4} & \mathbf{B}_{5,5} \end{bmatrix} \quad (37)$$

where  $\mathbf{B}_{i,j}$  is a  $17 \times 17$  sub-matrix that models how group  $j$  at time  $t$  contributes to group  $i$  at time  $t + 1$ . Note that  $\mathbf{B}_{3,1}$  is a zero matrix since women who are HIV- at time  $t$  cannot give birth to children who have been HIV positive for ten years by  $t + 1$  (i.e. five years into the future). Similar reasoning applies for the other zero submatrices.

The calculations involving  $\mathbf{B}_{1,j}$  produce the projection for the number of HIV- births (i.e.  $n_{1,1,t+1}$ ) contributed by duration group  $j$ . Similarly,  $\mathbf{B}_{2,j}$  projects the number of HIV positive births contributed by duration group  $j > 2$ .  $\mathbf{B}_{1,1}$  and  $\mathbf{B}_{2,1}$  are a little different in that they project each age group to the next oldest age group *and* from one HIV duration group to the next. Let us first consider  $\mathbf{B}_{1,1}$ :

$$\mathbf{B}_{1,1} = \begin{bmatrix} b_{1,1,t}^- & b_{2,1,t}^- & \cdots & & & b_{17,1,t}^- \\ p_{1,1,t}^- & 0 & \cdots & & & 0 \\ 0 & p_{2,1,t}^- & \ddots & & & \vdots \\ 0 & 0 & \ddots & & & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 0 & p_{16,1,t}^- & p_{17,1,t}^- \end{bmatrix}. \quad (38)$$

Recall that the number in the first age group at time  $t + 1$  is equal to the number of births summed across the fecund age groups. Let  $b_{a,d,t}^-$  be the factor needed to calculate the number

of HIV- births to mothers in age group  $a$  at time  $t$  and in duration group  $d$ :

$$b_{a,1,t}^- = s_{0,1,t} \frac{1}{1 + SRB} f_{a,1,t}^- \frac{1 + p_{a-1,1,t}^- \frac{n_{a-1,1,t}}{n_{a,1,t}}}{2}. \quad (39)$$

In our application of the multi-state, HIV-enabled CCMPP fertility only occurs among women aged 15-49, i.e.  $\alpha = 4$ ,  $\beta = 10$ . Consequently  $b_{a < 4,1,t}^- = b_{a > 10,1,t}^- = 0$ . In Equation 39 the factor  $\frac{1 + p_{a-1,1,t}^- \frac{n_{a-1,1,t}}{n_{a,1,t}}}{2}$  is used to approximate the number of women at risk of giving birth.<sup>4</sup> If the count in the denominator  $n_{a,1,t}$  is ever zero the entire ratio is simply replaced by zero. This issue arises when dealing with fertility of the HIV+ groups. The same procedure is used the analogous HIV+ equations if they involve dividing by zero. Recall that the  $p_{a-1,d,t}^-$  are the factors used to approximate the average number of women at the beginning and end of the age group.

$\mathbf{B}_{1,d}$  for  $d > 1$  projects forward HIV- births contributed by duration group  $d$  and can be written as:

$$\mathbf{B}_{1,d} = \begin{bmatrix} b_{1,d,t}^- & b_{2,d,t}^- & b_{3,d,t}^- & \cdots & b_{17,d,t}^- \\ 0 & \cdots & & & 0 \\ \vdots & \ddots & & & \\ 0 & & & & 0 \end{bmatrix} \quad (40)$$

where

$$b_{a,d,t}^- = s_{0,1,t} \frac{1}{1 + SRB} f_{a,d,t}^- \frac{1 + p_{a-1,d-1,t} \left( \frac{n_{a-1,d-1,t}}{n_{a,d,t}} \right)}{2} \quad (41)$$

for  $d > 1$ . The  $\mathbf{B}_{2,d}$ 's determine the number of people infected with HIV for less than five years at time  $t + 1$ , contributed by those in duration group  $d$  at time  $t$ . For the first two duration groups we have:

$$\mathbf{B}_{2,1} = \begin{bmatrix} b_{1,1,t}^+ & b_{2,1,t}^+ & \cdots & & b_{17,1,t}^+ \\ p_{1,1,t} & 0 & \cdots & & 0 \\ 0 & p_{2,1,t} & \ddots & & \vdots \\ 0 & 0 & \ddots & & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 0 & p_{16,1,t} & p_{17,1,t} \end{bmatrix} \quad (42)$$

The model does now allow HIV- women to become infected and give birth to HIV+ children in the same projection interval, zeros in the first rows. This assumption is justified by the low level of infectivity during the first five years after infection.

<sup>4</sup>In the matrix multiplication (i.e. the projection) this factor gets multiplied by the number of women in the appropriate age group,  $n_{a,1,t}$ . The  $p_{a-1,1,t}^-$  are defined in Equation ??.

$\mathbf{B}_{2,d}$  for  $d > 1$  projects forward the number of HIV+ births contributed by duration group  $d$ . It can be written as:

$$\mathbf{B}_{2,d} = \begin{bmatrix} b_{1,d,t}^+ & b_{2,d,t}^+ & b_{3,d,t}^+ & \cdots & b_{17,d,t}^+ \\ 0 & \cdots & & & 0 \\ \vdots & \ddots & & & \\ 0 & & & & 0 \end{bmatrix} \quad (43)$$

where

$$b_{a,d,t}^+ = s_{0,1,t} s_{0,1,t} \frac{1}{1 + SRB} f_{a,d,t}^+ \frac{1 + p_{a-1,d-1,t} \left( \frac{n_{a-1,d-1,t}}{n_{a,d,t}} \right)}{2} \quad (44)$$

The remaining non-zero sub-matrices –  $\mathbf{B}_{3,2}$ ,  $\mathbf{B}_{4,3}$ ,  $\mathbf{B}_{5,4}$  and  $\mathbf{B}_{5,5}$  – project people forward in both age and time, and consequently the only non-zero elements occur along the sub-diagonal:

$$\mathbf{B}_{i,j} = \begin{bmatrix} 0 & 0 & \cdots & & 0 \\ p_{1,d=j,t} & 0 & \cdots & & \vdots \\ 0 & p_{2,d=j,t} & \ddots & & \\ 0 & 0 & \ddots & & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 0 & p_{16,d=j,t} & p_{17,d=j,t} \end{bmatrix} \quad (45)$$

The Leslie matrix representation of the model greatly facilitates the implementation and use of the multi-state, HIV-enabled CCMPP using the R programming language (R Development Core Team, 2006). Repeated matrix multiplication produces the projected population at five-year intervals making it possible to explore the long term behavior of the population and the epidemic. See the results section for an application to a population living in the KwaZulu-Natal province of South Africa. Finally, if the Leslie matrix is irreducible and primitive then we can explore the stable age distribution of the population (Keyfitz and Caswell, 2005).

### 3 PARAMETER ESTIMATION

The CCMPP projections are used to estimate thirty-two of the model parameters. As mentioned earlier the vital rates, the initial population counts, and the HIV survival schedules are all fixed. The parameters we are interested in estimating are:

- $v$ : vertical transmission parameter that is constrained to be between 0 and 1; although the model is described as having a vertical transmission rate for each duration group, there are not enough data to estimate separate parameters – (1)

- $e_a$ : fertility selection parameter that is constrained to be equal to 1 for all groups except women aged 15-19 in the first HIV duration group, for whom we expect the value to be greater than 1 – (1)
- $g_d$ : fertility impairment parameter for women in duration group  $d$ , for  $d = 2, 3, 4$ ; the fertility impairment parameter for  $d = 5$  is constrained such that  $g_{d=5} = g_{d=4}$ , and the values for all duration groups are constrained to be between 0 and 1 – (3)
- $j_{a,k}$ : relative incidence rate ratio parameter that is constrained to be equal to 1 for women age 25-29 and non-negative for all other groups; values are estimated for women ( $k = 1$ ) age 15-19, 20-24, 30-34, 35-39, ..., 55-59 and for men in the age groups between 15-59, i.e. 8 and 9 age-specific parameters for women and men, respectively – (17)
- $H_h$ : scale parameters for the trend in HIV incidence for population  $h$  – (11)

For a given set of parameter values we obtain a set of not necessarily unique age- and sex-specific counts. These model outputs are used to calculate predicted values for the observed data. For example the ratio of the projected number of HIV+ women age 20-25 over the total number of women projected in that age group is used to predict the observed HIV prevalence for women in that age group. Several types of observed data, such as HIV prevalence, are used to estimate the values of the parameters that are most likely.

In this section we discuss this topic in greater detail. The first focus is on the types of data used in the analysis. We then shift to the likelihoods specific to each type of data. Finally we turn to the techniques used to estimate the parameters, namely maximum likelihood (ML) estimation and a bootstrapping approach.

### 3.1 Types of Data

Heuveline (2003) uses data published in the literature to estimate the model parameters. These data consist of observations from eleven different East African populations collected from antenatal clinics (ANCs), demographic surveillance sites, hospitals and general surveys. Both rural and urban areas are included, and the years of data collection range from 1989 to 1998. The data are classified into the following five categories (see Table 1, Heuveline, 2003):

1. HIV test results in a general-population sample (10 data sets)
2. HIV test results in an ANC-patient sample (3 data sets)
3. HIV test results in all or a sample of births from HIV+ mothers (3 data sets)
4. HIV test results during a follow-up of an HIV- sample (3 data sets)
5. Survival during a follow-up of HIV+ individuals (3 data sets)

The CCMPP outputs can be used to calculate predicted values for each type of data. These calculations and the likelihoods used to compare the predicted values to the data are addressed in the next section.

It should also be mentioned that the CCMPP also requires vital rates for the uninfected population and an initial age distribution. These model inputs are taken from the United Nations global population estimates (United Nations, 1999) and model life tables (United Nations, 1982).

## 3.2 Likelihoods

Each category of data provides the pieces needed for a proportion which leads to the use of the binomial distribution in the likelihood specification. The binomial likelihood can be written as:

$$\mathcal{L} = \prod \binom{N}{x} \pi^x (1 - \pi)^{N-x} \quad (46)$$

where  $N$  is the total number of events,  $x$  is the number of “successes”, and  $\pi$  is the probability of success. The first two quantities  $N$  and  $x$  are taken from the data and we use CCMPP to calculate  $\pi$ , given values for the parameters.

Before discussing the finer details of how the CCMPP outputs are used in the likelihood, it is important to cover two points. First, we need to temporally match up the CCMPP projections with the observed data. The start year for the model is the year when widespread transmission of HIV began,  $t_0$ . The population is then projected forward to the year when the data were collected. For example if widespread transmission in a country began in 1980 and the observed data are from 2000, then we can take the projected counts 20 years from the start time and compare these to the observed data. Given that the projections are in five year increments it is sometimes necessary to take the average across two projection periods to match the year of data collection. Estimates of when widespread HIV transmission began are taken from a report by the United Nations (1998, Table 1).

Second, the data come from populations at eleven different locations.<sup>5</sup> At a given location there can also be several different types of data. For example data from the population living in Mwanza, Tanzania include both HIV prevalence from a general population and survival information for those who are HIV+. As a result there is a separate likelihood for the 23 combinations of location and data type retrieved from the literature. These are indicated using  $h$  for the population (and location) and  $c$  for the type or category of data. Finally, the data consist of sex- and age-specific information so the likelihoods may also be indexed by these characteristics as well.

---

<sup>5</sup>Bujumbura, Burundi; Mangochi, Malawi; Mara, Tanzania; Mwanza, Tanzania; Fort Portal, Uganda; Gulu, Uganda; Masaka, Uganda; Rakai, Uganda; Lusaka, Zambia; Mposhi, Zambia and Mutasa, Zimbabwe.

### 3.2.1 HIV Test Results in a General-Population Sample

Various studies have collected data on sex- and age-specific HIV prevalence in a sample from the general population (Kilian et al., 1999; Nunn et al., 1997; Wawer et al., 1991; Serwadda et al., 1992; Shao et al., 1994; Grosskurth et al., 1995; Saidel et al., 1996; Fylkesnes et al., 1998; Gregson and Garnett, 2000). The age groups range from 15-19 to 55-59. This type of data, labeled ‘1’ ( $c = 1$ ), usually include the number of people tested and the percent who tested positive for HIV by sex and age. The likelihood, however, requires a count of individuals who are HIV+, so we calculate this quantity from the data and round it to the nearest integer. Let  $N_{a,k,t,h,c=1}$  denote the total number of individuals in age group  $a$  of sex  $k$  at time  $t$  at location  $h$  and let  $x_{a,k,t,h,c=1}$  be the number in this group who tested positive.

The projected counts from CCMPP  $n_{a,d,t,h}$  are used to predict sex- and age-specific prevalence for a given location as follows:

$$\pi_{a,k,t,h,c=1} = \frac{\sum_{d=2}^5 n_{a,d,t,h}}{\sum_{d=1}^5 n_{a,d,t,h}}, \quad (47)$$

where the sum is taken across HIV duration groups. Having chosen the projection period that matches up with the year the data were observed, we can use  $\pi_{a,k,t,h,c=1}$  in the binomial likelihood. A final note is that the observed data may be reported by age groups that do not match those of our projections. In this case weighted sums of the projected counts can be used to estimate HIV prevalence. For example if observed prevalence is reported for individuals age 22-30, then predicted prevalence can be calculated as:

$$\pi_{age=(22-30),k,t,h,c=1} = \frac{\sum_{d=2}^5 (n_{a=4,d,t,h} \times 0.6 + n_{a=5,d,t,h})}{\sum_{d=1}^5 (n_{a=4,d,t,h} \times 0.6 + n_{a=5,d,t,h})} \quad (48)$$

This issue arises with the other data types as well.

### 3.2.2 HIV Test Results in an ANC-Patient Sample

Four of the data sets used in this analysis provide age-specific information on HIV prevalence for female attendees of ANCs, typically from age 15 to 49 (Kilian et al., 1999; Fabiani et al., 2001; Slutsker et al., 1994; Sokal et al., 1993). This type of data, indexed by  $c = 2$ , takes a form similar to the observed prevalence from a general population except that they only refer to women. Both the total number of women tested  $N_{a,k=1,t,h,c=2}$  and the age-specific prevalence are reported. The data are included in the binomial likelihood as counts, so we calculate the number of women who tested positive  $x_{a,k=t,h}$  rounded to the nearest integer.

The predicted prevalence for the ANC attendees is calculated differently than for the general population. Recall that there are two primary assumptions of CCMPP concerning the fertility of HIV+ women. The first is that HIV+ women age 15-19 have higher fertility which is captured by the fertility selection parameter  $e_{a=4}$ . Second, fertility is expected to decline as time since infection increases; modelled by the fertility impairment parameters  $g_{d>1}$ . Since the HIV+ women observed in the data are pregnant, these parameters are included in the

calculation of predicted prevalence. The formula is:

$$\pi_{a,k=1,t,h,c=2} = \frac{\sum_{d=2}^5 n_{a,d,k=1,t,h} \times e_a \times g_d}{n_{a,d,k=1,t,h} + \sum_{d=2}^5 n_{a,d,k=1,t,h} \times e_a \times g_d}, \quad (49)$$

where the sum is taken over the duration groups. Having chosen the projection period that matches up with the year the data were observed, we can use  $\pi_{a,k=1,t,h,c=2}$  in the binomial likelihood.

### 3.2.3 HIV Test Results in all or a Sample of Births from HIV+ Mothers

Heuveline (2003) found three data sets consisting of information on the fertility of HIV+ women. However one of these sources Hira et al. (1989) differs from the others in that it provides information on whether or not an HIV+ mother infected her child. Data from the other two sources, Carpenter et al. (1997) and Gray et al. (1998), consist of the number of children born to both HIV+ and HIV- women, by age group. The likelihoods for the latter two sources are nearly identical to those in the data category  $c = 2$ , HIV test results in an ANC sample. The only difference is that the observed counts (i.e. the data) refer to the total number of children born to female ANC attendees in a specific age group, and the number of children born to HIV+ attendees. The probability that a child is born to an infected mother is calculated exactly the same as  $\pi_{a,k=1,t,h,c=2}$ . Given this similarity the data reported by Carpenter et al. (1997) and Gray et al. (1998) are classified here as  $c = 2$ .

Data that take the form of Hira et al. (1989) will be indicated by  $c = 3$ . The corresponding counts used in the likelihood refer to the total number of children born to infected mothers in age group  $o$ ,  $N_{o,t,h,c=3}$ , and the number of these children who are infected by their mothers  $x_{o,d=2,t,h,c=3}$ . The predicted rate of vertical transmission using the model outputs is calculated as:

$$\pi_{o,t,h,c=3} = \frac{\sum_{d=2}^5 n_{a,k=1,d,t,h} \times f_{a,d=1,t,h} \times e_a \times g_d \times v_d}{\sum_{d=2}^5 n_{a,k=1,d,t,h} \times f_{a,d=1,t,h} \times e_a \times g_d}, \quad (50)$$

where the sum is taken over the duration groups.

Unfortunately, this leaves only one data source to inform the estimate of the vertical transmission parameter. This issue is especially problematic when the parameter is constrained to be equal across duration groups (i.e.  $v_d = v$ , for all  $d$ ). Note that all the like terms in the numerator and denominator cancel out, so the projections have no influence on the likelihood. Thus including more data of this type would be beneficial for future analysis using this model.

### 3.2.4 HIV Test Results During a Follow-up of an HIV- Sample

Data on sex- and age-specific HIV incidence are also used to estimate CCMPP parameters. These data indexed by  $c = 4$  are typically reported in terms of the number of people who

become infected and the total number of person-years lived while uninfected (Kengeya-Kayondo et al., 1996; Wawer et al., 1994; Saidel et al., 1996). For the binomial likelihood however, we need the counts of the initial population observed  $N_{a,t,d=1,t,h,c=4}$  and the number of who become infected  $X_{a,t,d=2,t,h,c=4}$ . Thus the initial population size needs to be calculated from the observed data, and this calculation can be done as follows:

$$\text{Initial Population} = \frac{\# \text{ Converted}}{1 - \exp \left\{ -T \times \frac{\# \text{ Converted}}{\text{Person-Years}} \right\}} \quad (51)$$

where  $T$  is the total number of years observed.<sup>6</sup>

The model outputs are then used to calculate the probability of becoming infected for men or women in a certain age group at a given location and time. That is:

$$\pi_{a,k,d=1,t,h,c=4} = \frac{n_{a,k,d=1,t,h} \times s_{a,k,d=1,t,h} - n_{a+1,k,d=1,t+1,h}}{n_{a,k,d=1,t,h} \times s_{a,k,d=1,t,h}} \quad (52)$$

where  $\pi_{a,k,d=1,t,h,c=4}$  is the proportion of HIV- women/men who become infected after five years. As discussed earlier, the period of observation for the data may not be equal to the projection interval of five years. If the observation period is only four years, then the quantity of interest is calculated as:

$$1 - \exp \left\{ \frac{4}{5} \times \log \left( 1 - \frac{\# \text{ Converted}}{\text{Initial Population}} \right) \right\} \quad (53)$$

### 3.2.5 Survival During a Follow-up of HIV+ Individuals

The final category of data describes the survival (mortality) of HIV+ individuals presented by Nunn et al. (1997); Sewankambo et al. (1994); Todd et al. (1997). This category indexed by  $c = 5$  is similar to the previous one in that the data reported include the number of deaths observed among a cohort of HIV+ individuals of a particular age and sex  $X_{a,k,t,h,c=5}$  and the number of person-years observed for each group. As before, the likelihoods require the initial population size for each group  $N_{a,k,t,h,c=5}$  (see Equation 51). These two inputs  $X_{a,k,t,h,c=5}$  and  $N_{a,k,t,h,c=5}$  are the counts needed for the binomial likelihood with the corresponding proportion referring to the probability of death over a given period of time.

The procedure for calculating the probability of death from the model outputs is best described in two steps. First, we calculate the probabilities by age, sex, and duration group. This can be written as:

$$q_{a,k,d,t,h,c=5} = 1 - \left( \frac{n_{a+1,k,d+1,t+1,h}}{n_{a,k,d,t,h}} \right)^{\frac{T}{5}} \quad \text{for } d \geq 2 \quad (54)$$

---

<sup>6</sup>In deriving this equation it is helpful to note that the number of person-years lived by a population of initial size  $N_0$  and of size  $N_T$   $T$  years later is equal to  $\frac{(N_T - N_0) \times T}{\log(N_T/N_0)}$ .

where  $T$  is the number of years over which the data were observed. Since the observed data do not contain information on duration group, we must calculate the weighted average where the weights are the counts in each duration group. This second step is performed as follows:

$$\pi_{a,k,t,h,c=5} = \frac{\sum_{d=2}^5 q_{a,k,d,t,h,c=5} \times n_{a,k,d,t,h}}{n_{a,k,d,t,h}} \quad (55)$$

This is the probability used in the binomial likelihood.

### 3.3 Estimation Techniques

Maximum likelihood (ML) and bootstrapping techniques are used to estimate the most likely parameter values and the uncertainty around those point estimates. We begin by discussing the ML approach.

Given the data from an individual site, the likelihood of a specific set of CCMPP parameter values can be calculated using the binomial expressions described above. Twenty-two likelihoods are calculated, one for each of the twenty-two locations from which data come. We follow Heuveline (2003) and combine these likelihoods by taking the product across all twenty-two locations and data types, assuming independence. The set of parameter values that maximizes the combined likelihood is the ML point estimate.

The ML estimation as well as the implementation of CCMPP is performed using the R programming language (R Development Core Team, 2006). This languages provides an optimization routine `optim` that is used to find the parameter values that maximize the combined likelihood described above. In addition `optim` calculates the Hessian matrix of the likelihood function at the maximum. Standard errors are obtained by inverting the Hessian matrix and taking the square root of the diagonal elements.

A bootstrapping procedure is also used for estimation. We begin by sampling with replacement within four of the five data categories described earlier. Recall that there is only one data set providing information on the mother to child transmission of HIV, so this data set is included every time. Thus we are unable to produce an estimate of the vertical transmission parameter via this bootstrap technique. The ML estimates are then obtained using these “new” data, the estimates are stored, and the process is repeated over 300 times. The resulting distribution of ML estimates is used to assess possible values for the CCMPP parameters. Since the data sets are sampled with replacement some of the sites may not be included in the estimation making it impossible to estimate the corresponding site-specific parameters. Consequently the scale parameters  $H$  that indicate the size of the HIV epidemic for a given site are not estimated using the bootstrap approach.

## 4 RESULTS

The parameter estimates and 95% confidence intervals for the model parameters obtained by ML and bootstrap estimation are presented in Table 5 along with the original estimates

published by Heuveline (2003). Medians are used for the point estimates under the bootstrap approach. Confidence intervals for the parameters governing the level of the HIV epidemic for a given site were not reported by Heuveline. Lack of data and the characteristics of the bootstrap procedure, discussed above, prevent us from estimating these scale parameters and the vertical transmission parameters.

The point estimates for the parameters concerning both HIV and fertility are fairly similar across the three estimation methods – our ML and bootstrap methods and Heuveline’s ML method. Judging by the ML estimates, just under forty percent of children born to HIV+ mothers are likely to be infected. There is quite a bit of uncertainty around the point estimates, with the 95% confidence intervals stretching from 30% up to nearly 50%. This is rather high compared to the findings of other studies in Africa which usually fall in the range of 20-40% Working Group on Mother-To-Child Transmission (1995); Rollins et al. (2007). Women age 15-19 who are infected have fertility that is considerably higher than women in the same age group who are HIV-.<sup>7</sup> The point estimates for the fertility selection parameter vary, but each estimate is well within the confidence intervals of the other two. Our intervals, however, are much wider than those reported by Heuveline. This is also the case for the fertility impairment parameters, particularly for the bootstrap estimates. The interval for our ML estimate of the impairment parameter for women who have been infected for more than ten years includes negative values that are clearly outside the natural range for this parameter, and similarly the upper bound on the fertility impairment parameter for women infected for less than five years is greater than one, also a ‘nonphysical’ result as our colleagues in the hard sciences would term it. Fertility is expected to decline as time since infection increases, and both Heuveline’s and our ML estimates are consistent with this expectation, but all of the ML point estimates show a slight increase over the bootstrap results. However, the uncertainty around these estimates does not allow a definitive conclusion.

The point estimates of the relative incidence parameters for women are also similar across the three sets of estimates. Incidence among women peaks at ages 20-24 and declines at older ages. Again, we see more uncertainty around our estimates relative to Heuveline’s, with the ML confidence interval including negative (and thus nonphysical) values. Another difference is that our estimates for incidence of women ages 15-19 fall outside the 95% confidence interval reported by Heuveline. Finally, uncertainty around the estimates increases with age for each set of results which makes sense because there are fewer data describing HIV incidence and prevalence at older ages.

Compared to women, incidence among men is estimated to peak at older ages. Heuveline’s results show a peak in the late twenties while our results put the peak during the early thirties. However, given the uncertainty in these estimates we are unable to definitively pinpoint the peak. Aside from the later peak among men, the results are similar to those for women. Some of our confidence intervals using ML techniques include nonphysical values, and there is generally more uncertainty around our results at older ages.

---

<sup>7</sup>The multiplicative factor by which fertility is higher among infected women is equal to the fertility selection factor times the fertility impairment parameter of women infected less than five years. These factors are 1.42, 1.58, and 1.58 for Heuveline’s analysis, our ML estimates and the bootstrap estimates, respectively. This modification to fertility only applies while women are in the states of being within the ages of 15 to 19, and being HIV+ for 0 to 4 years.

Table 5: Estimates of CCMPP Parameters and 95% Confidence Intervals.

<b>Parameters</b>	<b>Heuveline</b>	<b>ML</b>	<b>Bootstrap</b>
Vertical Transmission	0.385 (0.297, 0.478)	0.395 (0.303, 0.487)	-- --
Fertility Selection	1.672 (1.492, 1.865)	1.836 (1.331, 2.341)	1.754 (1.28, 2.367)
Fertility Impairment duration 0-4	0.848 (0.798, 0.909)	0.86 (0.671, 1.049)	0.899 (0.452, 1)
duration 5-9	0.357 (0.276, 0.450)	0.356 (0.067, 0.645)	0.251 (0.013, 0.954)
Duration 10+	0.293 (0.078, 0.607)	0.125 (-0.322, 0.572)	0.291 (0, 0.901)
Female Relative Incidence Ratio			
15-9	0.594 (0.545, 0.650)	0.511 (0.414, 0.609)	0.525 (0.435, 0.691)
20-4	1.325 (1.239, 1.412)	1.291 (1.051, 1.531)	1.294 (1.139, 1.45)
25-9	1.00 --	1.00 --	1.00 --
30-4	0.752 (0.647, 0.886)	0.797 (0.578, 1.016)	0.783 (0.6, 0.913)
35-9	0.635 (0.482, 0.762)	0.533 (0.338, 0.727)	0.533 (0.447, 0.727)
40-4	0.551 (0.409, 0.795)	0.579 (0.326, 0.832)	0.557 (0.375, 0.76)
45-9	0.356 (0.159, 0.544)	0.385 (0.124, 0.647)	0.394 (0.201, 0.621)
50-4	0.295 (0.095, 0.679)	0.281 (-0.045, 0.607)	0.277 (0.015, 0.836)
55-9	0.246 (0.087, 0.627)	0.431 (0.035, 0.828)	0.393 (0, 1.214)
Male Relative Incidence Ratio			
15-9	0.059 (0.024, 0.109)	0.104 (0.061, 0.147)	0.101 (0.038, 0.186)
20-4	0.583 (0.483, 0.684)	0.477 (0.366, 0.588)	0.488 (0.302, 0.755)
25-9	1.149 (0.986, 1.285)	0.999 (0.796, 1.203)	1.011 (0.868, 1.213)
30-4	0.936 (0.773, 1.130)	1.021 (0.771, 1.272)	1.014 (0.843, 1.228)
35-9	0.759 (0.573, 0.944)	0.716 (0.454, 0.977)	0.721 (0.569, 0.945)
40-4	0.769 (0.554, 1.007)	0.769 (0.463, 1.075)	0.779 (0.438, 1.104)
45-9	0.622 (0.409, 0.879)	0.679 (0.344, 1.015)	0.643 (0.328, 0.967)
50-4	0.417 (0.120, 0.773)	0.234 (-0.132, 0.601)	0.266 (0.013, 0.896)
55-9	0.168 (0.001, 0.445)	0.189 (-0.098, 0.476)	0.231 (0, 0.924)
Level Of HIV Epidemic			
Bujumbura, Burundi	0.627 --	0.59 (0.485, 0.695)	-- --
Mangochi, Malawi	0.253 --	0.252 (0.202, 0.303)	-- --
Mara, Tanzania	0.224 --	0.234 (0.192, 0.276)	-- --
Mwanza, Tanzania	0.137 --	0.143 (0.121, 0.166)	-- --
Fort Portal, Uganda	0.565 --	0.81 (0.666, 0.954)	-- --
Gulu, Uganda	0.717 --	0.75 (0.611, 0.889)	-- --
Masaka, Uganda	0.347 --	0.361 (0.304, 0.419)	-- --
Rakai, Uganda	0.619 --	0.656 (0.536, 0.777)	-- --
Chelston, Zambia	0.94 --	1.088 (0.92, 1.257)	-- --
Kapira Mposhi, Zambia	0.41 --	1.135 (0.899, 1.371)	-- --
Mutasa, Zimbabwe	0.501 --	0.752 (0.639, 0.865)	-- --

Table 6: Comparison of Model Fit to Data on Age-Specific HIV Prevalence from Kapiri Mposhi, Zambia.

<b>WOMEN</b>			
Age Groups	Data <sup>a</sup>	Heuveline	Thomas & Clark
15–9	0.082	0.05	0.132
20–4	0.246	0.111	0.28
25–9	0.177	0.091	0.233
30–9	0.198	0.064	0.168
40–9	0.178	0.042	0.113
Sum of Squared Errors		0.063	0.012

<b>MEN</b>			
Age Groups	Data <sup>a</sup>	Heuveline	Thomas & Clark
15–9	0.055	0.005	0.014
20–4	0.058	0.049	0.13
25–9	0.209	0.097	0.248
30–9	0.319	0.078	0.201
40–9	0.123	0.063	0.164
Sum of Squared Errors		0.077	0.024

<sup>a</sup>Source: Fylkesnes et al. (1998)

We now turn to the parameters that describe the scale of the HIV epidemic at a given site. There are several differences between our estimates and those presented by Heuveline (2003). We estimate that the epidemics are much larger in Fort Portal, Uganda, Kapira Mposhi, Zambia, and Mutasa, Zimbabwe. Of these, the most troubling is the Zambian site. Our point estimate is more the double that of Heuveline’s (1.135 vs. 0.41). This difference is explored by calculating the predicted values for the data from Kapiri Mposhi, Zambia using both our estimate of the level parameter and Heuveline’s; all of the other parameters are set to Heuveline’s estimates. The results are presented in Table 6 for both women (upper panel) and men (lower panel). The observed age-specific prevalence is shown in the column labeled “Data”, and Heuveline’s and our predicted values or shown in the third and fourth columns, respectively. Each panel also includes the sum of squared errors (SSE) by sex for each set of estimates. Based on this measure our predicted values fit better than those of Heuveline for both women (SSE: 0.012 vs. 0.63) and men (SSE: 0.024 vs. 0.077). Similar results hold for both Fort Portal, Uganda and Mutasa, Zimbabwe.<sup>8</sup>

The estimates presented above are obtained using data from countries located in East Africa. This gives rise to the question of regional variation within sub-Saharan Africa. We explore this question with an application of CCMPP to a population living in a rural area of the KwaZulu-Natal province of South Africa. Data from this population published by Welz et al.

<sup>8</sup>The sum of squared errors for the data from Fort Portal, Uganda are 0.014 (ours) vs. 0.037 (Heuveline) for women and 0.001 (ours) vs. 0.27 (Heuveline) for men. The corresponding numbers for the data from Mutasa, Zimbabwe are 0.002 vs. 0.041 for women and 0.010 vs. 0.053 for men.

(2007) include age-specific HIV prevalence for the following age groups: 15-19, . . . , 45-49 for women; and 15-19, . . . , 50-54 for men. The data are further classified by resident status with non-residents being individuals whose main residence is elsewhere but who maintain regular contact with the surveyed household through periodic visits. Prevalence is generally much higher among non-residents which is consistent with the finding that migrants are at a higher risk of becoming infected. We model this difference by estimating CCMPP with separate parameters for the level of the epidemic for residents and non-residents. The results from this model are presented in Table 7 under the column labeled full model. We also report estimates for a reduced model for which incidence is constrained to be equal across certain age groups, for women 30-49 and for men 25-34 and 35-54.

Relative to males, incidence among females is much higher in the youngest age group. Incidence peaks among women between the ages of 20 and 29 while the peak for men occurs during the late twenties and early thirties. There is a relatively large drop in relative incidence after the peak for women. Recall that women between the ages of 25 and 29 serve as the reference group, with their relative incidence fixed at a value of 1. Relative incidence for women in the next oldest age group has a 95% confidence interval of (-0.11, 0.51). There is a similar finding for men, but it is much less pronounced. With respect to the two resident groups, residents and non-residents, the estimates do not indicate a difference in the level of the epidemic.

The amount of uncertainty around the estimates in the full model motivated the estimation of a more parsimonious model. Note how the confidence intervals become increasingly wider with age<sup>9</sup>. With relatively few data we are unable to identify differences between the point estimates for relative incidence after the peak. The ‘reduced model’ column in Table 7 displays estimates for a simpler model that collapses all post-peak age groups for women and includes only two parameters for relative incidence among men older than 34. The results show an age pattern similar to that seen for the full model, and there is more precision around the estimates for this reduced model. We again find little support for the hypothesis that the HIV epidemic is at a higher level among non-residents, but in both cases the estimates of epidemic scale are much higher than those from the sites in East Africa.

An interesting comparison between the full and reduced models is to project prevalence and combine the age groups to obtain an estimate of adult prevalence which can then be compared to observed prevalence.<sup>10</sup> To get a sense of the uncertainty around these estimates we appeal to asymptotics and sampled CCMPP inputs from a multivariate normal distribution with the mean and covariance matrix taken from the ML estimates. For each set of inputs in the sample we run CCMPP and the corresponding outputs are used as a predictive distribution. The 0.025 and 0.975 quantiles of this predictive distribution are reported together as the 95% predictive interval (95% PI).

Welz et al. (2007) report that 21.5% of all adult residents aged 15-49 are HIV+ (prevalence for non-residents is not given). Using the ML estimates as the model inputs gives a predicted

---

<sup>9</sup>It should also be pointed out that two of the confidence intervals include negative values that are problematic since these parameters are naturally bounded at zero.

<sup>10</sup>This partially serves as a check for the age distribution of the initial population and the choice of vital rates.

Table 7: CCMPP Parameter Estimates and 95% Confidence Intervals using Data from KwaZulu-Natal, South Africa.

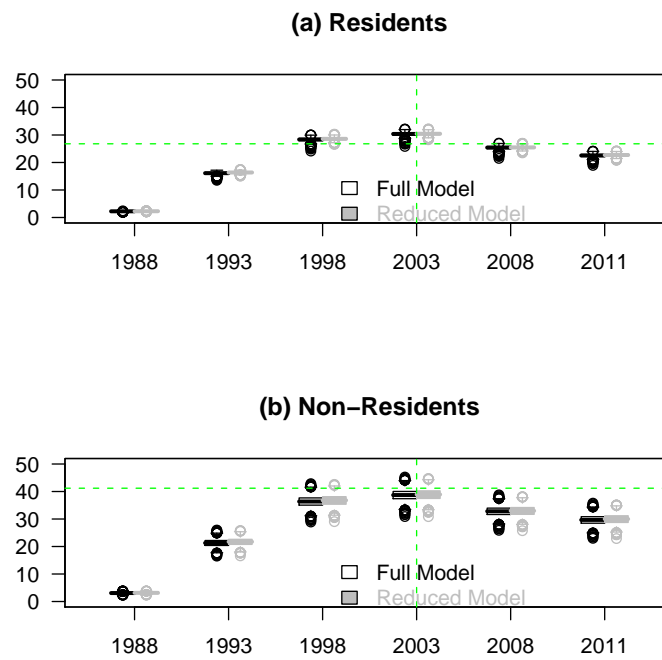
<b>PARAMETERS</b>	<b>FULL MODEL</b>	<b>REDUCED MODEL</b>
Female Relative Incidence Ratio		
15-9	0.26	0.27
	(0.2, 0.33)	(0.22, 0.32)
20-4	0.63	0.64
	(0.43, 0.82)	(0.49, 0.79)
25-9	1	1
	–	–
30-4	0.2	0.29 <sup>a</sup>
	(-0.11, 0.51)	(0.24, 0.34)
35-39	0.4	
	(0.15, 0.64)	
40-44	0.25	
	(0.02, 0.49)	
45-49	0.21	
	(0.04, 0.37)	
Male Relative Incidence Ratio		
15-19	0.02	0.02
	(0.01, 0.03)	(0.01, 0.03)
20-24	0.2	0.21
	(0.14, 0.27)	(0.16, 0.27)
25-29	0.69	0.65 <sup>b</sup>
	(0.47, 0.91)	(0.54, 0.77)
30-34	0.58	
	(0.28, 0.88)	
35-39	0.22	0.28 <sup>c</sup>
	(-0.12, 0.56)	(0.23, 0.34)
40-44	0.4	
	(0.05, 0.74)	
45-49	0.18	
	(-0.09, 0.45)	
50-54	0.32	
	(0.09, 0.55)	
Scale		
Residents	1.77	1.73
	(1.41, 2.13)	(1.5, 1.95)
Non-Residents	2.45	2.39
	(1.86, 3.03)	(1.95, 2.83)

<sup>a</sup>This is the parameter estimate for women between the ages of 30-49.

<sup>b</sup>This is the parameter estimate for men between the ages of 25-34.

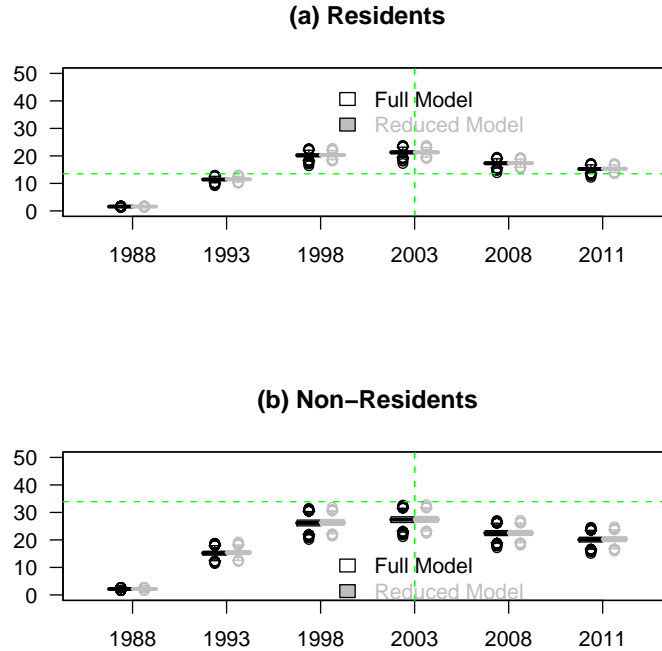
<sup>c</sup>This is the parameter estimate for men between the ages of 35-54.

Figure 2: Projected Prevalence Over Time for Women Aged 15-49.



Note: The horizontal and vertical dashed lines indicate the prevalence and year, respectively, from the observed data.

Figure 3: Projected Prevalence Over Time for Men Aged 15-54.

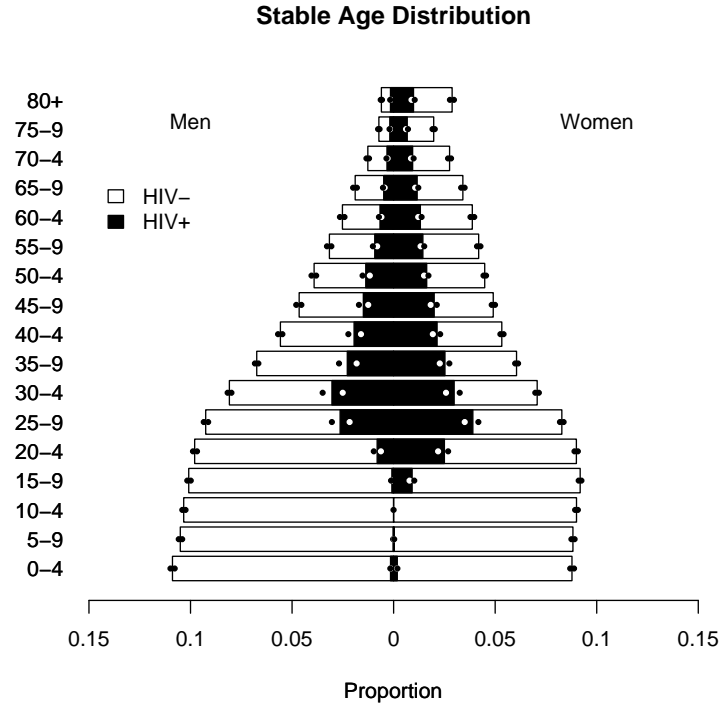


Note: The horizontal and vertical dashed lines indicate the prevalence and year, respectively, from the observed data.

prevalence of 25.9% (95% PI: 24.8 – 26.9%) which is much too narrow. Figures 2 & 3 show the projected prevalence for women and men, respectively. There is a pair of boxplots for each five-year increment over time up to 2003, the year of data collection, and ten years into the future. The boxplot on the left (right) refers to the full (reduced) model. Among women the projections match the observed prevalence of adults much better for non-residents than for residents. Predicted prevalence is too high for adult female residents. As seen in Panel (a) of Figure 2 the predictive interval for the full model barely includes the observed prevalence in 2003, while the corresponding interval for the reduced model is too narrow. Among non-resident women (Panel (b)), for whom prevalence is much higher, the predictive intervals for both models cover the observed prevalence in 2003. Finally the predictive intervals tend to be larger for non-residents in both the full and reduced models.

The results for men are shown in Figure 3 with residents presented in Panel (a) and non-residents in Panel (b). The findings are similar to those of women in that the projections tend to be too high for residents and too low for non-residents. The fit appears to be slightly worse among men in that the predictive intervals appear to be too narrow. As found with women, the reduced model provides very similar projections to the more complicated model.

Figure 4: Projected Population Pyramids at 25 Year Intervals for KwaZulu-Natal South Africa.



Note: The white bars indicate the proportion of the population in each age group (HIV- and HIV+ combined), while the black bars indicate the proportion of the total population who are HIV+ in each age group. 95% confidence intervals are depicted by the dots.

Given the high level of observed prevalence in KwaZulu-Natal it is interesting to explore the long term implications of the estimated CCMPP parameters for the population. The eigenvector corresponding to the largest real eigenvalue of the Leslie matrix is the stable age distribution (Keyfitz and Caswell, 2005). Population pyramids for the stable equivalent populations are presented in Figure 4 with men on the left and women on the right. The white horizontal bars represents the proportion of the total population in an age group and the black section of each bar indicates the proportion of the population in that age group that is HIV+. The dots at the end of each bar indicate the 95% confidence intervals around the proportion represented by the bar – they are very narrow.

In the stable population HIV prevalence resembles the age profile for incidence. There is an earlier peak for women occurring among women between the ages of 25 and 29. Among men the peak is at the next older age group, and HIV prevalence is generally lower relative to women. Prevalence is so high among women (25%) that the female stable population is actually shrinking. In contrast the male stable population (prevalence 17%) is growing slowly – which hints at possible pressure on the sex ratio of this population. This ‘stable equivalent’

analysis reveals a potential problem with the current parameterization of the model, revealed by the proportion older than 65 that is infected. In the absence of widespread antiretroviral therapy that increases survival with HIV it is unlikely that a large proportion of the elderly population is infected. Since only those younger than 60 are able to be infected in this model we expect few to survive to ages much beyond 65. This result may suggest that we are using survival rates for the HIV+ population that are too high.

## 5 DISCUSSION

In this article we extend the work of Heuveline (2003) by developing a Leslie matrix representation of his multi-state, HIV-enabled CCMPP and providing new estimates for countries in East Africa and for a rural population living in the KwaZulu-Natal province of South Africa. Our findings are broadly similar to Heuveline's in that for women incidence peaks between the ages of 20 and 29 and for men between the ages of 25 and 34, and this is true for both East Africa and the small population KwaZulu-Natal. Our results also include estimates for the level of the epidemic at each site from which the data come. The findings reveal substantial variation across East Africa, and the point estimates suggest a much larger epidemic in KwaZulu-Natal relative to the sites in East Africa.

Focusing on the estimation method we compare the maximum likelihood estimates of the model parameters to those obtained via a bootstrapping procedure. The results raise some questions concerning the reliability of the maximum likelihood estimates. We find that several confidence intervals for the CCMPP parameters include negative values which is troubling since all of the parameters are bounded below by zero. Further the bootstrap analysis yields much wider confidence intervals, suggesting the ML results understate the uncertainty around the parameter estimates.<sup>11</sup> An alternative approach to estimation is the use of Bayesian methods. Poole and Raftery (2000) developed the Bayesian melding technique specifically for use with deterministic models like CCMPP, see also Alkema et al. (2007).

A Bayesian framework is particularly appealing given the potential need for model comparison, such as the full and reduced models fit to the South African data. Recall that for the KwaZulu-Natal model the projections for adult prevalence were generally too low for non-residents and too high for residents. Given that prevalence is higher among women, it may be inappropriate to model the epidemic's level for men and women simultaneously since non-resident men and resident women have similar levels of prevalence. This suggests another model with level parameters for each combination of sex and resident status, perhaps with addition variation related to the full and reduced specifications shown in Table 7. In a situation like this it would be invaluable to have an estimation approach that objectively compares or even combines model fits, and Bayesian approaches exist for both, see Raftery (1995).

The Leslie matrix is very useful in that it facilitates the actual implementation of the model

---

<sup>11</sup>Recall that the predictive intervals for adult prevalence at the KwaZulu-Natal site were too narrow for resident men and women as well as non-resident men.

and allows us to explore the stable characteristics of a population experiencing an HIV epidemic. The results suggest that the survival schedules for HIV+ individuals produce unusually large proportions of HIV+ individuals at older ages. This finding may be plausible with future improvements in the access to health care and the development of new medications to extend life with HIV, but the survival schedules were not originally designed to capture these potential changes.<sup>12</sup> Since it is possible that future health-related innovations and improvements may take a long time to develop, it would be helpful to have a set of survival rates informed by data from the populations of interest. Another alternative would be to include these rates among the list of estimated parameters.

Our current and future work on this model is addressing the limitations and possible improvements discussed above. Most important, we are working on a fully Bayesian implementation of the estimation procedure and updating the empirical database to include as much contemporary data as possible. The resulting estimates will be more robust and describe the current state of contemporary HIV epidemics across the African continent.

## REFERENCES

- Alkema, L., Raftery, A. E., and Clark, S. J. 2007. “Probabilistic Projections of HIV Prevalence Using Bayesian Melding”. *The Annals of Applied Statistics* 1(1):229–248.
- Bowley, A. 1924. “Births and Population in Great Britain”. *Economic Journal* 34(134):188–92.
- Cannan, E. 1895. “The Probability of a Cessation of the Growth of Population in England and Wales During the Next Century”. *Economic Journal* 5(20):505–15.
- Carpenter, L. M., Nakiyingi, J. S., Ruberantwari, A., Malamba, S. S., Kamali, A., and Whitworth, J. A. 1997. “Estimates of the Impact of HIV Infection on Fertility in a Rural Ugandan Population Cohort”. *Health Transition Review* 7(Supplement 2):113–126.
- Chin, J. 1994. “A Beginner’s Guide for Understanding and Using the Epimodel – Version 2: A Computer Program for Estimation and Projection of AIDS Cases and Deaths”. School of Public Health, University of California, Berkeley.
- Chin, J. and Lawnga, S. 1991. “Estimation and Projection of Adult AIDS Cases: A Simple Epidemiological Model”. *Bulletin of the World Health Organization* 69(4):399–406.
- Day, J. C. 1996. “Population Projections of the United States by Age, Sex, Race, and Hispanic Origin: 1995 to 2050”. Current Population Reports P25-1130, U.S. Bureau of the Census, U.S. Government Printing Office, Washington, DC.
- Dorn, H. F. 1950. “Pitfalls in Population Forecasts and Projections”. *Journal of the American Statistical Association* 45(251):311–334.

---

<sup>12</sup>It would be more satisfying to model this process explicitly by including additional states to CCMPP for individuals how are on antiretroviral therapies, for example.

- Fabiani, M., Accorsi, S., Lukwiya, M., Rosolen, T., Ayella, E., Onek, P., and Declich, S. 2001. "Trend in HIV-1 Prevalence in an Antenatal Clinic in North Uganda and Adjusted Rates for the General Population". *AIDS* 15:97–103.
- Ford, K. and Hosegood, V. 2005. "AIDS mortality and the mobility of children in KwaZulu Natal, South Africa". *Demography* 42(4):757–768.
- Fylkesnes, K., Ndhlovu, Z., Dasumba, D., Musonda, R., and Sichone, M. 1998. "Studying the Dynamics of the HIV Epidemic: Population-Based Data Compared with Sentinel Surveillance in Zambia." *AIDS* 12:1227–34.
- Garenne, M. L., Tollman, S. M., Collinson, M. A., and Kahn, K. 2007. "Fertility trends and net reproduction in Agincourt, rural South Africa, 1992-2004". *Scandinavian Journal of Public Health* 35:68–76.
- Ghys, P. D., Brown, T., Grassly, N. C., Garnett, G., Stannecki, K. A., Stover, J., and Walker, N. 2004. "The UNAIDS Estimation and Projection Package: a software package to estimate and project national HIV epidemics". *Sexually Transmitted Infections* 80:i5–i9.
- Gray, R. H., Wawer, M. J., Serwadda, D., Sewankambo, N., Li, C., Wabwire-Mangen, F., Paxton, L., Kiwanuka, N., Kigozi, G., Konde-Lule, J., Quinn, T. C., and Gaydos, C. A. 1998. "Population-Based Study of Fertility in Women with HIV-1 Infection in Uganda". *Lancet* 351(9096):98–103.
- Gregson, S. and Garnett, G. 2000. "Contrasting Gender Differentials in HIV-1 Prevalence and Associated Mortality Increase in Eastern and Southern Africa: Artefact of Data or Natural Course of Epidemics?" *AIDS* 14(Suppl.3):S85–99.
- Gregson, S., Nyamukapa, C., Lopman, B., Mushati, P., Garnett, G. P., Chandiwana, S. K., and Anderson, R. M. 2007. "Critique of early models of the demographic impact of HIV/AIDS in sub-Saharan Africa based on contemporary empirical data from Zimbabwe". *Proceedings of the National Academy of Sciences of the United States of America* 104(37):14586–14591.
- Grosskurth, H., Mosha, F., Todd, J., Senkoro, K., Newell, J., Lkoko, A., Changalucha, J., West, B., Mayaud, P., Gavyole, A., Gabone, R., Mabey, D., and Hayes, R. 1995. "A Community Trial of the Impact of Improved Sexually Transmitted Disease Treatment on the HIV Epidemic in Rural Tanzania." *AIDS* 9:927–34.
- Heuveline, P. 2003. "HIV and Population Dynamics: A General Model and Maximum-Likelihood Standards for East Africa". *Demography* 40(2):217–245.
- Hira, S., Kamanga, J., Bhat, G., Mwale, C., Tembo, G., Luo, N., and Perine, P. 1989. "Perinatal Transmission of HIV-1 in Zambia". *British Medical Journal* 299(6718):1250–52.
- Hunter, S. C., Isingo, R., Boerma, J. T., Urassa, M., Mwaluko, G. M. P., and Zaba, B. 2003. "The association between HIV and fertility in a cohort study in rural Tanzania". *Journal of Biosocial Science* 35(2):189–199.

- Kahn, K., Tollman, S. M., Collinson, M. A., Clark, S. J., Twine, R., Clark, B. D., Shabangu, M., Gomez-Olive, F. X., Mokoena, O., and Garenne, M. L. 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* 35:8–20.
- Kengeya-Kayondo, J., Kamali, A., Nunn, A., Ruberantwari, A., Wagner, H., and Mulder, D. 1996. “Incidence of HIV-1 infection in Adults and Socio-Demographic Characteristics of Seroconverters in a Rural Population in Uganda: 1990-1994”. *International Journal of Epidemiology* 25:1077–82.
- Keyfitz, N. and Caswell, H. 2005. *Applied Mathematical Demograph.* Third edition edition. Springer.
- Kilian, A., Gregson, S., Ndyanabangi, B., Walusaga, K., Kipp, W., Sahlmüller, G., Garnett, G., Asiimwe-Okiror, G., Kabagambe, G., Weis, P., and von Sonnenburg, F. 1999. “Reductions in Risk Behaviour Provide the Most Consistent Explanation of Declining HIV Prevalence in Uganda.” *AIDS* 13:391–98.
- Leslie, P. 1945. “On the Use of Matrices in Certain Population Mathematics”. *Biometrika* 33(3):183–212.
- Lewis, J. J., Ronsmans, C., Ezeh, A., and Gregson, S. 2004. “The Population Impact of HIV on Fertility in sub-Saharan Africa”. *AIDS* .
- Morgan, D., Mahe, C., Mayanja, B., Okongo, J., Lubega, R., and Whitworth, J. 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* 16:597–603.
- Nunn, A., Mulder, D., Kamali, A., Ruberantwari, A., Kengeya-Kayondo, J., and Whitworth, J. 1997. “Mortality Associated with HIV-1 Infection Over Five Years in a Rural Ugandan Population: Cohort Study”. *British Medical Journal* 315(7111):767–71.
- Nyirenda, M., Hosegood, V., Barnighausen, T., and Newell, M.-L. 2007. “Mortality Levels and Trends by HIV Serostatus in Rural South Africa”. *AIDS* 21(suppl 6):S73–S79.
- Pearl, R. and Reed, L. J. 1920. “On the Rate of Growth of the Population of the United States Since 1790 and its Mathematical Representation”. *Proceedings of the National Academy of Science* 6(6):275–288.
- Poole, D. and Raftery, A. E. 2000. “Inference for Deterministic Simulation Models: The Bayesian Melding Approach”. *Journal of the American Statistical Association* 95(452):1244–1255.
- Pritchett, H. 1891. “A Formula for Predicting the Population of the United States”. *Publications of the American Statistical Association* 2(14):278–286.
- R Development Core Team. 2006. *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. 1995. "Bayesian Model Selection in Social Research". *Sociological Methodology* 25:111–163.
- Rollins, N., Little, K., Mzolo, S., Horwood, C., and Newell, M.-L. 2007. "Surveillance of mother-to-child transmission prevention programmes at immunization clinics: the case for universal screening". *AIDS* 21:1341–1347.
- Saidel, T., Sokal, D., Rice, J., Buzingo, T., and Hassig, S. 1996. "Validation of a Method to Estimate Age-Specific Human Immunodeficiency Virus (HIV) Incidence in Developing Countries Using Population-Based Seroprevalence Data". *American Journal of Epidemiology* 144:214–23.
- Serwadda, D., Wawer, M., Musgrave, S., Sewankambo, N., Kaplan, J., and Gray, R. 1992. "HIV Risk Factors in Three Geographic Strata of Rural Rakai District, Uganda." *AIDS* 6:983–89.
- Sewankambo, N., Wawer, M., Gray, R., Serwadda, D., Li, C., Stallings, R., Musgrave, S., and Konde-Lule, J. 1994. "Demographic Impact of HIV Infection in Rural Rakai District, Uganda: Results of a Population-Based Cohort Study". *AIDS* 8:1707–13.
- Shao, J., Burbaker, G., Levin, A., Kibauri, A., Massesa, E., Siso, Z., Konings, E., Clayton, Y., Kumby, D., Alexander, S., Waters, D., Drummond, J., Biggar, R., Scott, G., Miller, G., Goedert, J., and Blattner, W. 1994. "Population-Based Study of HIV-1 Infection in 4,086 Subjects in Northwest Tanzania". *Journal of Acquired Immune Deficiency Syndrome* 7:397–402.
- Slutsker, L., Cabeza, J., Wirima, J., and Steketee, R. 1994. "HIV-1 Infection Among Women of Reproductive Age in a Rural District in Malawi". *AIDS* 8:1337–40.
- Sokal, D., Buzingo, T., Nitunga, N., Kadende, P., and Standaert, B. 1993. "Geographic and Temporal Stability of HIV Seroprevalence Among Pregnant Women in Bujumbura, Burundi". *AIDS* 7:1481–84.
- Terceira, N., Gregson, S., Zaba, B., and Mason, P. R. 2003. "The Contribution of HIV to Fertility Decline in rural Zimbabwe, 1985-2000". *Population Studies* 57(2):149–164.
- Timaes, I. M. and Jasseh, M. 2004. "Adult mortality in sub-Saharan Africa: Evidence from demographic and health surveys". *Demography* 41(4):757–772.
- Todd, J., Balira, R., Grosskurth, H., Mayaud, P., Mosha, F., ka-Gina, G., Klokke, A., Gabbone, R., Gavyole, A., Mabey, D., and Hayes, R. 1997. "HIV-Associated Adult Mortality in a Rural Tanzanian Population". *AIDS* 11:801–07.
- United Nations. 1982. *Model Life Tables for Developing Countries*. New York: United Nations.
- United Nations. 1998. "AIDS, Mortality, and Population Change – Technical Meeting on the Demographic Impact of HIV/AIDS, Tuesday, 10 November 1998". Available online at <http://www.un.org/esa/population/pubsarchive/hivmtg/aidsrep.pdf> (downloaded on February 13, 2008).

- United Nations. 1999. *World Population Prospects: the 1998 Revision*. New York: United Nations.
- United Nations. 2004. *World Population Prospects: the 2002 Revision*, volume 3. New York: United Nations.
- Wachter, K. W., Knodel, J. E., and Vanlandingham, M. 2002. "AIDS and the elderly of Thailand: Projecting familial impacts". *Demography* 39(1):25–41.
- Wawer, M., Serwadda, D., Musgrave, S., Konde-Lule, J., Musagara, M., and Sewankambo, N. 1991. "Dynamics of Spread of HIV-1 Infection in a Rural District of Uganda". *British Medical Journal* 303(6813):1303–06.
- Wawer, M., Sewankambo, N., Berkley, S., Serwadda, D., Musgrave, S., Gray, R., Musagara, M., Stallings, R., and Konde-Lule, J. 1994. "Incidence of HIV-1 Infection in a Rural Region of Uganda". *British Medical Journal* 303(6813):1303–06.
- Welz, T., Hosegood, V., Jaffar, S., Batzing-Feigenbaum, J., Herbst, K., and Newell, M. 2007. "Continued Very High Prevalence on HIV Infection in Rural Kwazulu-Natal, South Africa: A Population-Based Longitudinal Study". *AIDS* 21:1467–1472.
- Whelpton, P. 1936. "An Empirical Method of Calculating Future Population". *Journal of the American Statistical Association* 31(195):457–473.
- Working Group on Mother-To-Child Transmission. 1995. "Rates of mother-to-child transmission of HIV-1 in Arica, America, and Europe: Results from 13 Perinatal Studies". *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* 8:506–510.
- Zaba, B. and Gregson, S. 1998. "Measuring the impact of HIV on fertility in Africa". *Aids* 12:S41–S50.
- Zaba, B., Marston, M., Crampin, A. C., Isingo, R., Biraro, S., Barnighausen, T., Lopman, B., Lutalo, T., Glynn, J. R., and Todd, J. 2007. "Age-specific mortality patterns in HIV-infected individuals: a comparative analysis of African community study data". *Aids* 21:S87–S96.