Probabilistic Projections of Populations With HIV: A Bayesian Melding Approach

Samuel J. Clark and Jason Thomas[1]

ABSTRACT

Population projection models are valuable tools for demographers and public policy makers alike. A particular example is the model developed by Heuveline (2003) which captures some of the links between population growth and the spread of HIV/AIDS. This model requires relatively few inputs and can provide projections of HIV prevalence for populations for which reliable data are limited. We reproduce Heuveline's work, but in a Bayesian context. More specifically, we use Bayesian melding to obtain measures of uncertainty around both the model inputs and outputs in the form of probability distributions. This approach provides useful information to policy makers concerning issues with planning and resource allocation.

[1]Samuel J. Clark, Department of Sociology, University of Washington, 202 Savery Hall, Box 353340, Seattle, WA 98195-3340; E-mail: samclark@u.washington.edu. Jason Thomas, Department of Sociology, University of Washington; E-mail: method@u.washington.edu. This research was supported by the Center for Statistics and the Social Sciences (CSSS) at the University of Washington. The authors are grateful to Patrick Heuveline for his correspondence and for making his work available. We are also indebted to Leontine Alkema for sharing her code for implementing Bayesian melding for the Estimation and Projection Package.

INTRODUCTION

Many governments and policy organizations rely on population projections to help them adequately provide services for the population. For some populations, generating realistic projections is seriously complicated by the lack of data and the complex dynamics of HIV/AIDS epidemics experienced by the population. An important step in confronting these difficulties has been made by Heuveline (2003), who developed a model which captures several of the dependencies inherent in the relationship between population growth and the spread of HIV/AIDS. Heuveline has also published default values for most of the model parameters, which are estimated using data from several East African countries. This allows age- and sex-specific projections, by HIV infections status, to be made from a small amount of data.

The purposes for this article are twofold; the first of which is to facilitate the use of Heuveline's model. The model, referred to here as CCMPP, is an extension of the cohort component method for making age- and sex-specific population projections (for an introduction to the cohort component method see Preston et al., 2001; Keyfitz and Caswell, 2005). Heuveline has extended the basic version of the cohort component method by expanding the state space through which individuals can make transitions. In total, there are five states: (1) HIV–, (2) having been infected with HIV for up to four years (3) HIV+ for five to nine years, (4) HIV+ for 10-5 years, and (5) HIV+ for more than 15 years. As the population is projected through time, the transitions that individuals make are determined by aging, (age- and sex-specific) HIV incidence rates, and being born with or without HIV. This last process is the feature of CCMPP which captures two of the connections between the dynamics of

populations and HIV/AIDS epidemics. More specifically, CCMPP models the vertical transmission of HIV from mother to child, as well as the negative relationship between fertility and duration with HIV. Sexual activity is expected to be higher among young women (i.e. those ages 15-9) who are HIV+, which consequently increases fertility. However, as the time since infection increases, fecundability – and thus fertility – both decrease.

Although CCMPP requires over 20 inputs, Heuveline (2003) provides default values for most of the parameters, which are applicable to populations located in East Africa. Thus, only two model inputs are needed, i.e. the year in which the HIV epidemic began and an estimate of HIV prevalence, to produce age- and sex-specific projections of HIV prevalence. The ability of the model to produce such valuable outputs (generated from HIV and population dynamics) is the justification for why we wish to facilitate the use of CCMPP. To help achieve this objective, CCMPP is presented here using matrix notation, which provides the guidelines for an efficient implementation of the model using any standard programming language. Furthermore, we are developing a package for the `R` programming language R Development Core Team (2006) which provides several functions for running CCMPP (for a given set of inputs) and analyzing the outputs. This package will be made publicly available from the authors upon completion.

The second purpose for this paper is to provide alternative estimates of uncertainty around the model's inputs and outputs. Heuveline (2003) reports confidence intervals for the CCMPP inputs, which can be used to generate (age-specific) probabilistic projections of HIV prevalence. We propose a different approach, namely Bayesian melding, which is specifically designed for deterministic models (Raftery et al., 1995; Poole and Raftery, 2000). Bayesian

melding provides a way to estimate various sources of uncertainty around the model inputs and outputs. This technique is applied to CCMPP to obtain (posterior) distributions for the parameter inputs, as well as probabilistic projections of HIV prevalence . These results are compared to those reported by (Heuveline, 2003) obtained from maximum likelihood estimation.

The paper is organized as follows. First, we describe CCMPP and present the model using matrix notation. This is followed by a discussion of the parameter estimation via maximum likelihood. Third, we review Bayesian melding and the application to CCMPP. Then we present the results and conclude with a discussion of possible extensions for CCMPP and of ways in which Bayesian methods can enhance the use of CCMPP even further.

MODEL

Single-State CCMPP

CCMPP is a deterministic model for discrete time that produces age- and sex-specific projections. Age groups are advanced over the projection interval[1] by applying age-specific survivorship ratios. CCMPP can also include migration, but our presentation of the model will only describe a closed population. For all age groups, excluding the youngest and oldest, the projection can be written as

[1]The projection interval is typically set equal to the length of the age groups, excluding the oldest group which is open-ended.

$$n_{a+1,t+1} \;=\; s_{a,t}\; n_{a,t} \tag{1}$$

where $n_{a,t}$ is the number of women in age group a at time t, and $s_{a,t}$ is the survivorship ratio, or the proportion of women in age group a that survives to age group $a+1$ from time t to $t+1$. The projection for the oldest age group, which is open-ended, is calculated as

$$n_{a',t+1} \;=\; s_{a'-1,t}\; n_{a'-1,t} \;+\; s_{a',t}\; n_{a',t} \tag{2}$$

where a' is the oldest age group. The youngest age group is projected forward by applying the appropriate survival ratio to the total number of female births. The latter quantity is calculated by applying the age-specific fertility rate to the average number of women alive at the beginning and end of the projection interval[2] and summing over the age groups. This is written as

$$n_{1,t+1} \;=\; \sum_{a=\alpha}^{\beta} s_{0,t}\; \frac{1}{1+SRB}\; f_{a,t}\; \frac{(1+s_{a,t})}{2}\; n_{a,t} \tag{3}$$

where $f_{a,t}$ is the fertility rate for women in age group a at time t and SRB is the ratio of male to female births, or the sex ratio at birth,[3] and the lower and upper bounds of the childbearing age range are α and β, respectively.

The preceding equations can be efficiently represented using matrix notation. To illustrate, consider a population with three age groups. We can write the CCMPP as

[2]This quantity is used to approximate the number of person-years lived by women during the projection interval.

[3]The number of male births is calculated by multiplying the total number of births by $\frac{SRB}{1+SRB}$.

$\mathbf{n_{t+1}} = \mathbf{A_t n_t}$, where $\mathbf{n}_t$ is a 3×1 column vector containing the counts in each age group at time t, and $\mathbf{A_t}$ is the so-called *Leslie* matrix. For this population, the *Leslie* matrix is

$$\mathbf{A_t} = \begin{bmatrix} b_{1,t} & b_{2,t} & b_{3,t} \\ s_{1,t} & 0 & 0 \\ 0 & s_{2,t} & s_{3,t} \end{bmatrix} \tag{4}$$

where

$$b_{a,t} = s_{0,t} \frac{1}{1+SRB} f_{a,t} \frac{(1+s_{a,t})}{2}. \tag{5}$$

Note that when $a < \alpha$ or $a > \beta$ we have $f_{a,t} = 0$, and thus $b_{a,t} = 0$.

Multi-State CCMPP

Heuveline (2003) has developed CCMPP for a population with 5 different groups, distinguished by HIV status. There are four HIV duration groups (i.e. 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 5 HIV groups. The model is introduced with a series of equations representing the transition from one group/time period to the next. This is followed by a presentation of the model in the form of a Leslie matrix.

Begin by dividing the population into age groups, where $a = 1, 2, \ldots, 17$ correspond to age groups $0-4, 5-9, \ldots, 80+$. Denote membership in the HIV duration groups by d, with $d = 1, 2, \ldots, 5$ corresponding to HIV–, HIV+ for 0-4 years, ..., HIV+ for more than 15

years. The time period is indexed by t, but note that the length between t and $t+1$ is equal to the width of a standard age interval, i.e. 5 years. Let $n_{a,d,t}$ be the number of women in age group a and duration group d at time t. For $17 > a > 1$, we have

$$n_{a+1,1,t+1} = n_{a,1,t}\ s_{a,1,t}\ (1 - i_a) \tag{6}$$

$$n_{a+1,2,t+1} = n_{a,1,t}\ s_{a,1,t}\ i_a\ s_{a,2,t} \tag{7}$$

$$n_{a+1,d,t+1} = n_{a,d-1,t}\ s_{a,1,t}\ s_{a,d,t}, \quad \text{for } d > 2 \tag{8}$$

where $s_{a,d,t}$ is the survivorship ratio for age group a and duration group d at time t. Note that for $5 > d > 2$ the 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 survival rate as the HIV– group, as well as an additional force of mortality specific to d. 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

$$i_{a,t} = 1 - \exp\left\{-\Gamma_{t-t_0}\ H\ j_a\right\} \tag{9}$$

where Γ_{t-t_0} is the parametric curve used to model the trend in the HIV epidemic, which depends on the time since the epidemic began (t_0). The parameter H is a population-specific scale parameter which captures the size of the epidemic. The parameter j_a is the age- and sex-specific scaling factor of incidence relative to women aged 20-25, for whom the parameter is constrained to be one in order for the model to be identifiable (i.e. $j_5 = 1$).

These equations are slightly different for the youngest and oldest age groups. The oldest (open-ended) age group is incremented by 2 sources: those aged 75-9 and 80+ in the previous time period. This gives us (for $a = 17$)

$$\begin{aligned}
n_{17,1,t+1} &= n_{16,1,t}\, s_{16,1,t}\, (1 - i_{16}) \\
&\quad + n_{17,1,t}\, s_{17,1,t}\, (1 - i_{17}), && (10) \\
n_{17,2,t+1} &= n_{16,1,t}\, s_{16,1,t}\, i_{16}\, s_{16,2,t} \\
&\quad + n_{17,1,t}\, s_{17,1,t}\, i_{17}\, s_{17,2,t}, && (11) \\
n_{17,d,t+1} &= n_{16,d-1,t}\, s_{16,1,t}\, s_{16,d,t} \\
&\quad + n_{17,d-1,t}\, s_{17,1,t}\, s_{17,d,t} \quad \text{for } 2 < d < 5, && (12) \\
n_{17,5,t+1} &= n_{16,4,t}\, s_{16,1,t}\, s_{16,5,t} + n_{17,4,t}\, s_{17,1,t}\, s_{17,5,t} \\
&\quad + n_{16,5,t}\, s_{16,1,t}\, s_{16,5,t} + n_{17,5,t}\, s_{17,1,t}\, s_{17,5,t}. && (13)
\end{aligned}$$

As seen with the single-state CCMPP, the first age group is projected forward by applying age-specific fertility rates to the average number of women (at the beginning and end of the projection interval) who are in the corresponding age and HIV duration groups. Heuveline's model also has three additional parameters included in the fertility calculations. These parameters model a few of the connections between HIV and population dynamics. First, consider the number of HIV– births

$$n_{1,1,t+1} = s_{0,1,t}\, \frac{1}{1 + SRB} \times$$

$$\left(\sum_{a=\alpha}^{\beta} f_{a,1,t} \frac{n_{a,1,t} + p^{-}_{a-1,1,t} n_{a-1,1,t}}{2} \; + \; \sum_{d=2}^{5} \sum_{a=\alpha}^{\beta} f^{-}_{a,d,t} \frac{n_{a,d,t} + p_{a-1,d-1,t} n_{a-1,d-1,t}}{2} \right). \quad (14)$$

In the equation 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 α and β. Fertility among HIV+ women introduces the following parameters

$$f^{-}_{a,d,t} \;=\; f_{a,1,t}\; e_a\; g_d\; (1 - v_d) \quad (15)$$

for $d > 1$. The superscript in $f^{-}_{a,d,t}$ designates births to children 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, or the vertical transmission rate. The parameter e_a captures the higher level of sexual activity and resulting fertility among HIV+ women aged 15-9 who have been infected for 0-4 years ($d = 2$). In other words, we expect $e_{a=4} > 1$, and $e_{a\neq=4}$ are constrained to be one. The parameter g_d represents the fertility impairment for women in duration group d which is expected to become stronger as the time since infection increases. The corresponding equations for children who are HIV+ are

$$n_{1,2,t+1} \;=\; s_{0,1,t}\; \frac{1}{1 + SRB} \sum_{d=2}^{5} \sum_{a=\alpha}^{\beta} f^{+}_{a,d,t}\; \frac{n_{a,d,t} + p_{a-1,d-1,t}\; n_{a-1,d-1,t}}{2}, \quad (16)$$

$$f^{+}_{a,d,t} \;=\; f_{a,1,t}\; e_a\; g_d\; v_d. \quad (17)$$

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 $p_{a,d,t}$. These parameters can be written as

$$p^{-}_{a,1,t} \;=\; s_{a,1,t}\; (1 - i_a) \quad (18)$$

$$p_{a,1,t} = s_{a,1,t}\, i_a\, s_{a,2,t} \tag{19}$$

$$p_{a,d,t} = s_{a,1,t}\, s_{a,d,t}, \quad \text{for } d > 1. \tag{20}$$

These equations for the multi-state CCMPP can conveniently be expressed in matrix notation. For a population with 17 age groups and 5 HIV duration groups, the population at time t is represented by a (85×1) column vector

$$\mathbf{n}_t = \left[\begin{array}{c} n_{1,1,t} \\ n_{2,1,t} \\ \vdots \\ n_{17,1,t} \\ \hline \vdots \\ \hline n_{1,4,t} \\ n_{2,4,t} \\ \vdots \\ n_{17,4,t} \end{array}\right]$$

The corresponding *Leslie* matrix is

$$\mathbf{A}_t = \begin{bmatrix} \mathbf{B}_{1,1} & \mathbf{B}_{1,2} & \mathbf{B}_{1,3} & \mathbf{B}_{1,4} & \mathbf{B}_{1,5} \\ \mathbf{B}_{2,1} & \mathbf{B}_{2,2} & \mathbf{B}_{2,3} & \mathbf{B}_{2,4} & \mathbf{B}_{2,5} \\ \mathbf{0} & \mathbf{B}_{3,2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_{4,3} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_{5,4} & \mathbf{B}_{5,5} \end{bmatrix} \tag{21}$$

where $\mathbf{B}_{i,j}$ is a (17×17) submatrix which 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 matrices.

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 highest. Let us first consider $\mathbf{B}_{1,1}$

$$\mathbf{B}_{1,1} = \begin{bmatrix} b^-_{1,1,t} & b^-_{2,1,t} & \cdots & & & b^-_{17,1,t} \\ p^-_{1,1,t} & 0 & \cdots & & & 0 \\ 0 & p^-_{2,1,t} & \ddots & & & \vdots \\ 0 & 0 & \ddots & & & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 0 & p^-_{16,1,t} & p^-_{17,1,t} \end{bmatrix}.$$

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

$$b^-_{a,1,t} \;=\; s_{0,1,t}\,\frac{1}{1+SRB}\,f^-_{a,1,t}\,\frac{1+p^-_{a-1,1,t}\frac{n_{a-1,1,t}}{n_{a,1,t}}}{2}. \tag{22}$$

In our application of the multi-state CCMPP fertility only occurs among women aged 15 to 49 (i.e. $\alpha = 4,\ \beta = 10$). Therefore, $b^-_{a<4,1,t} \;=\; b^-_{a>10,1,t} \;=\; 0$. In the equation above, the factor $\frac{n_{a-1,1,t}}{n_{a,1,t}}$ is used to approximate the number of women at risk of giving birth. If the count in the denominator, i.e. $n_{a,1,t}$, is ever zero, simply replace the entire ratio by zero. This issue arises when dealing with fertility of the HIV+ groups. The same procedure is used for those equations if they involve dividing by zero.

$\mathbf{B}_{1,d}$ for $d > 1$ projects HIV– births contributed by the duration group d. It can be written as

$$\mathbf{B}_{1,d} = \begin{bmatrix} b^-_{1,d,t} & b^-_{2,d,t} & b^-_{3,d,t} & \cdots & b^-_{17,d,t} \\ 0 & \cdots & & & 0 \\ \vdots & \ddots & & & \\ & & & & \\ 0 & & & & 0 \end{bmatrix} \tag{23}$$

where

$$b^-_{a,d,t} \;=\; s_{0,1,t}\,\frac{1}{1+SRB}\,f^-_{a,d,t}\,\frac{1+p_{a-1,d-1,t}(\frac{n_{a-1,d-1,t}}{n_{a,d,t}})}{2} \tag{24}$$

(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

$$\mathbf{B}_{2,1} = \begin{bmatrix} b^+_{1,1,t} & b^+_{2,1,t} & \cdots & & & b^+_{17,1,t} \\ p_{1,1,t} & 0 & \cdots & & & 0 \\ 0 & p_{2,1,t} & \ddots & & & \vdots \\ 0 & 0 & \ddots & & & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 0 & p_{16,1,t} & p_{17,1,t} \end{bmatrix}.$$

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

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

$$\mathbf{B}_{2,d} = \begin{bmatrix} b^+_{1,d,t} & b^+_{2,d,t} & b^+_{3,d,t} & \cdots & b^+_{17,d,t} \\ 0 & \cdots & & & 0 \\ \vdots & \ddots & & & \\ & & & & \\ & & & & \\ 0 & & & & 0 \end{bmatrix} \tag{25}$$

where

$$b^+_{a,d,t} \;=\; s_{0,1,t}\; s_{0,1,t}\; \frac{1}{1+SRB}\; f^+_{a,d,t}\; \frac{1 + p_{a-1,d-1,t}(\frac{n_{a-1,d-1,t}}{n_{a,d,t}})}{2}. \tag{26}$$

The remaining non-zero submatrices, i.e. $\mathbf{B}_{3,2}, \mathbf{B}_{4,3}, \mathbf{B}_{5,4}, \mathbf{B}_{5,5}$, project each age group to the next oldest, and from one duration group to the next. Thus, the only non-zero elements occur along the subdiagonal

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

Bayesian Melding with CCMPP

The characteristic which probably distinguishes the Bayesian approach most is the treatment of unknown parameters as random variables. The parameters we wish to make inferences for, θ, are first quantified as a probability density, $p(\theta)$, which characterizes prior beliefs about the parameters. These prior beliefs are then updated using observed data, $\mathbf{y}$. This process is carried out by specifying a conditional probability of observing the data for given values of the parameters, $L(\mathbf{y}|\theta)$, also known as the likelihood. Bayes' Theorem is used to update the prior as

$$p(\theta|\mathbf{y}) = \frac{L(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} \tag{27}$$

$$\propto L(\mathbf{y}|\theta)p(\theta) \tag{28}$$

and inferences for θ are made using the posterior distribution, $p(\theta|\mathbf{y})$. The second line in the equations above refers to the fact that $p(\mathbf{y})$ does not depend on θ, so the posterior distribution only needs to be known up to a constant.

Bayesian melding applies this estimation strategy to situations in which a deterministic model, such as CCMPP, is used in the likelihood component. Let M represent the model

which transforms a set of parameter inputs, θ, into a set of outputs, $\phi = M(\theta)$. The 22 model inputs for CCMPP include the parameters relating to sex- and age-specific HIV incidence (j_a), fertility selection of women aged 15-9 (e_4), fertility impairment (g_d), and vertical transmission (v_d). To carry out the estimation procedure, we specify a prior density for these inputs, $p(\theta)$, and a likelihood for the outputs and the data, $L(M(\theta))$ (more details to follow). These two sources of information are combined to produce the following posterior distribution for the model inputs:

$$p(\theta|\mathbf{y}) \propto L(\mathbf{y}|M(\theta))p(\theta).$$

Inference is performed by sampling from $p(\theta|\mathbf{y})$ and summarizing the sample information. Furthermore, we can run CCMPP for each set of inputs sampled, which generates a sample from the posterior of model outputs, $p(\phi|\mathbf{y})$. A version of the Sampling Importance Resampling (SIR) algorithm can be used to sample from the posterior (Rubin, 1987; Poole and Raftery, 2000). The steps are as follows

1. Draw a sample $\{\theta_{\mathbf{1}}, \theta_{\mathbf{2}}, \ldots, \theta_{\mathbf{n}}\}$ from the prior distribution $p(\theta)$.

2. Calculate $\{\phi_{\mathbf{1}}, \phi_{\mathbf{2}}, \ldots, \phi_{\mathbf{n}}\}$ by running CCMPP for each $\theta_{\mathbf{i}}$.

3. Calculate the sampling importance weights:

$$w_i = \frac{L(\mathbf{y}|M(\theta_{\mathbf{i}}))}{\sum_{i=1}^{n} L(\mathbf{y}|M(\theta_{\mathbf{i}}))}$$

4. Draw a sample from $\{\theta_{\mathbf{1}}, \theta_{\mathbf{2}}, \ldots, \theta_{\mathbf{n}}\}$ using w_i as sampling weights. This serves as an approximation to the posterior of the inputs. Run CCMPP for the resampled θ to obtain a sample from the posterior of the outputs.

Priors. The specification of the prior density for the CCMPP inputs reflects our high level of uncertainty in the parameter values. We drew a sample of size 500,000 with half coming from diffuse uniform distributions and the other half coming from normal distributions centered on each parameter's maximum likelihood estimate (mle)[4]. This mixture ensures that the probability space is thoroughly explored while still giving adequate weight to regions of high likelihood. Thus, the prior density is a multivariate distribution containing 22 independent random variables (one for each input).

The endpoints of the uniform distribution that the vertical transmission parameter is drawn from are zero and one, which reflect the natural boundaries of a proportion. Although we expect the fertility impairment parameters to fall between these same boundaries, these are partially drawn from uniform distributions ranging from zero to two (with the other half being drawn from a normal distribution centered around the mle). This provides the estimation procedure with some robustness to our assumption being wrong, again reflecting the uncertainty around the parameter values. Half of the sample of fertility selection parameters is drawn from a uniform ranging from zero to 5. Most of the uniform distributions chosen for the incidence parameters range from zero to two. These parameters model the age profile of HIV incidence. *A priori,* we believe these profiles take on concave shapes (i.e. does not hold water), but we are uncertain where the peak occurs. Therefore, the range on the uniform distribution for men (women) ranges from zero to three for those aged 25 to 39(20 to 29). All of these values are listed in Table .

[4]If the mle for a given parameter is close to zero, then we draw the sample from a truncated normal with a cutoff at zero.

Likelihood. The binomial distribution is used to calculate the likelihood of the 22 CCMPP parameters (j_a, e_a, g_d, v_d,) and the data. The data used to evaluate the likelihood are taken from 23 studies of 11 East African populations.[5] Each data source provides observations on at least one of the following outcomes: (1) HIV test results in a general-population sample, (2) HIV test results in an ANC-patient sample, (3) HIV test results in all or a sample of births from HIV+ mothers, (4) HIV tests results during a follow-up of an HIV– sample, and (5) survival during a follow-up of HIV+ individuals (see Table 1, Heuveline, 2003). These outcomes are converted into proportions, which are used with the binomial distribution to estimate the parameters of interest.

Consider an example using HIV test results to help describe the estimation procedure. First, we use the data to calculate the proportion of respondents in a particular age group who are HIV+, namely $\pi_{a,t}$. The population from which these data are collected is then modeled using CCMPP. In order to project the population of interest, we need information on the survival and fertility rates for HIV– individuals, as well as the year in which the epidemic began for this population (i.e. $s_{a,d,t}, f_{a,d,t}, t_0$). Data from the United Nations (1998, 1999) supply the values needed for these model inputs, which are treated as fixed. Given a set of values of the parameters in which we are interested in estimating, CCMPP produces age- and sex-specific projections starting from the onset year of the epidemic. In the third step, we

[5]Most of the data are not random samples from the respective population. The geographic regions from which these data come from include Fort Portal, Uganda; Gulu, Uganda; Masaka, Uganda; Rakai, Uganda; Mara, Tanzania; Mwanza, Tanzania; Bujumbura, Burundi; Mangochi, Malawi; Lusaka, Zambia; Mposhi, Zambia; and Mutasa; Zimbabwe.

match the year in which the data are collected to the closest projection year. For example, if the data are from 1998 and the epidemic in this particular country began in 1987, then we would take the second projection period. The projected counts from this period are used to calculate the total number of people, $N_{a,t}$, as well as the total number of HIV+ individuals, $n_{a,+,t}$ in the age group. We can now use the binomial distribution to calculate the likelihood of these projections given the observed data

$$\binom{N_{a,t}}{n_{a,+,t}}(\pi_{a,t})^{n_{a,+,t}}(1-\pi_{a,t})^{N_{a,t}-n_{a,+,t}}. \tag{29}$$

This procedure is performed for each of the data sets to produce 23 separate likelihoods. Taking the natural logarithm of the likelihoods and summing gives us a total log likelihood that combines the information across the studies. The total log likelihood can be thought of as a function of j_a, e_a, g_d, v_d,. The set of values for these model inputs which maximizes the total log likelihood is the set of ML estimates. We use the non-linear minimization routine, `nlm`, in the `stats4` package of the R programming language to maximize the total log likelihood.[6] This routine also returns the Hessian matrix which is used to calculate standard errors.

In the original analysis, Heuveline finds that the model with the best fit to the data includes 3 duration groups for the fertility impairment parameter (g_d), 1 vertical transmission parameter (v_d), and parameters for the relative incidence ratios for the age groups from 15-9 to 55-9. Altogether there are 22 parameters estimated.[7] Figure 1 presents Heuveline's

[6]The R programming language is also used to implement CCMPP; the code is available upon request.

[7]Recall that the fertility selection parameter, e_a, is constrained to equals 1 for each age

estimates along with our replication of his analysis. The dots in the plot represent the point estimates, and the lines indicate the standard errors (see Table 2 in the Appendix for the actual values). There are a few noticeable differences between the two sets of results. Our estimates of the relative incidence ratios for females ages 20-4 and 30-4 appear to be higher, while our estimate of the fertility impairment parameter for the first HIV duration group seems to be lower. The 95% confidence intervals from our results are generally tighter as well. However, the results are fairly similar overall and we feel confident that our implementation of CCMPP is working.

RESULTS

Histograms of the posterior samples for the CCMPP inputs are presented in Figures 3 – 5. In these plots, the smooth black line indicates the prior distribution, the dashed, grey line depicts the ML estimate, and the black, horizontal line represents the median of the posterior sample. These histograms are helpful, but we prefer to summarize the posteriors with 95% confidence intervals shown in Figure 2, which are compared to the mle results.

Perhaps the most striking difference between these two estimation procedures is that Bayesian melding results suggest much more variation in the parameter values. Virtually all of the Bayesian intervals are wider, particularly for the CCMPP parameters relating to the older age groups. This finding makes intuitive sense because there are very few data on the HIV incidence of men and women over the age of 40. Thus, we would expect a lot of variation in these distributions. This is also the case for the fertility impairment of HIV+

group except 15-9; and the relative incidence ratio, j_a, is also fixed at 1 for females ages 25-9.

women who have been infected for more than ten years. Again, there are few observations of women in this state, which brings up the question of why the frequentist intervals are so tight. Heuveline's estimates are much wider than our intervals as well, which may point to a problem with mle procedure.

The Bayesian melding procedure also provides a means for including uncertainty around the CCMPP outputs. This is accomplished by running the model for each posterior sample and calculating the output of interest. To illustrate, we ran CCMPP for each of the 11 populations included in the ML estimation. Age-specific HIV prevalence was then calculated from these projections (pooled across all of the populations). The results are presented in Figure 6. Each boxplot summarizes the range of outputs for a particular age group. The red line imposed over the boxplots indicates the projected prevalence using the ML estimates as the CCMPP inputs.

There is an increase in the median posterior prevalence for each age group for the 3 periods after the onset of the epidemic. After 15 years, however, the changes are not consistent across the age groups. Given the fact that the uncertainty in the posterior prevalence also increases over time it is difficult to make strong claims about relative changes. However, it is interesting to note that the median posterior prevalence declines among women ages 15-9, but continues to increase for women ages 25 to 49. Another point worth mentioning is that the ML estimates tend to fall near the lower end of the posterior prevalence distributions. This is driven by the early finding that several of the posterior samples for the relative incidence ratios resemble the flat prior distributions, while the ML estimates tend toward zero.

It is also possible to generate uncertainty around age- and sex-specific prevalence using only the results from the ML estimation. If we assume that the ML estimates of the CCMPP inputs follow a normal distribution (and are independent) with the mean and variance equal to the ML estimate and the squared, standard error (used to calculate the confidence interval), then we can sample from these distributions and run CCMPP for each sample. These steps were taken with the results reported by Heuveline (2003), and used to calculate projections of age-specific HIV prevalence. These results are presented in Figure 7.

Compared to the Bayesian melding results, there is much less uncertainty around the projected prevalence using only the information from the ML procedure. As expected, the level of projected prevalence is also much lower compared to the earlier results. The final point to be made is that the current technique produces nonsensical projections (i.e. negative prevalence) for certain samples from the asymptotic distributions of the ML estimates. These projections can be ignored or truncated to zero prevalence, but neither of these options provides a satisfactory solution to the problem since they discard the variation indicated by the estimates.

DISCUSSION

In this analysis we applied Bayesian melding to a deterministic projection model. This estimation procedure adequately models the various sources of uncertainty around both the model inputs and outputs. Furthermore, this technique can be used with other methods to inform sources of uncertainty not addressed in this analysis. For example, given the limited data at the older age ranges, a model which collapses the older age groups may provide more

reliable estimates. Bayesian melding can be carried out in a broader analysis which looks at model selection or even model averaging. This is a very promising direction for future research in this area.

REFERENCES

Heuveline, P. 2003. "HIV and Population Dynamics: A General Model and Maximum-Likelihood Standards for East Africa". *Demography* 40(2):217–245.

Keyfitz, N. and Caswell, H. 2005. *Applied Mathematical Demograph*. Third edition edition. Springer.

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.

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

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., Givens, G. H., and Zeh, J. E. 1995. "Inference from a Deterministic Population Dynamics Model for Bowhead Whales (with discussion)". *Journal of the American Statistical Association* 90:402–416.

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.

United Nations. 1998. "AIDS, Mortality, and Population Change – Technical Meeting on the Demographic Impact of HIV/AIDS, Tuesday, 10 November 1998". Available on-line at http://www.undp.org/popin/popdiv/hivmtg/aidshiv1.htm (downloaded on February 16, 1999).

United Nations. 1999. *World Population Prospects: the 1998 Revision*. New York: United Nations.

FIGURES

Figure 1: Maximum Likelihood Estimates of the CCMPP Parameters with 95% Confidence Intervals (indicated by lines)

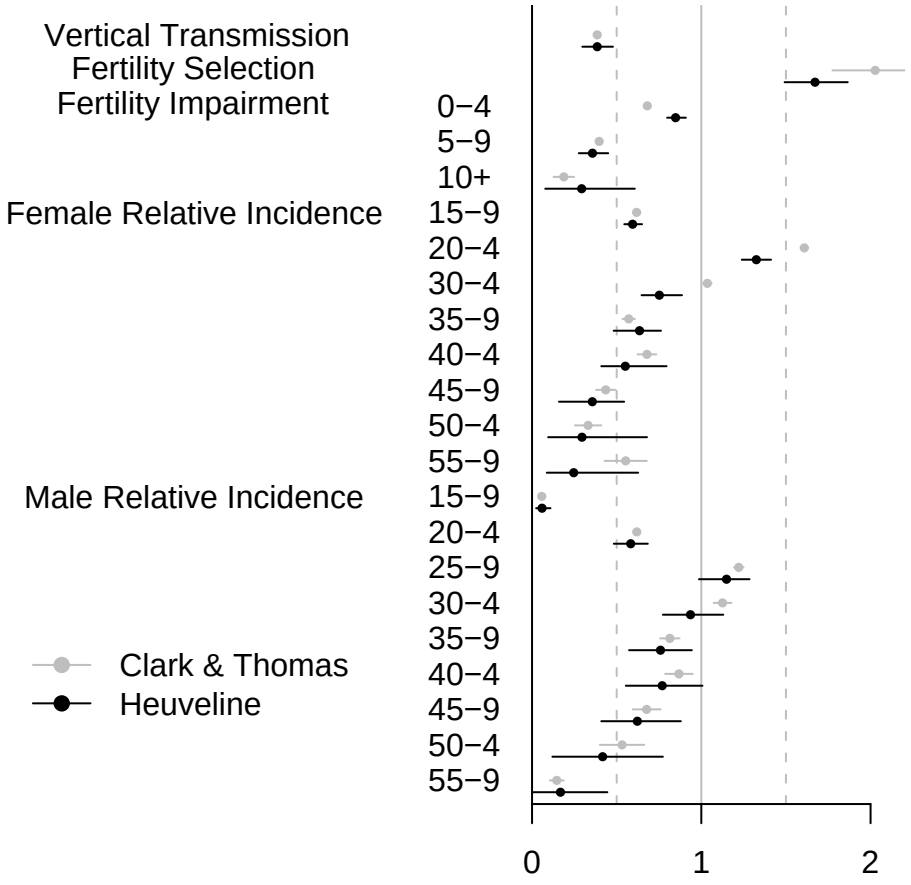


Figure 2: 95% Confidence Intervals of the CCMPP Parameters from Maximum Likelihood Estimation and Bayesian Melding

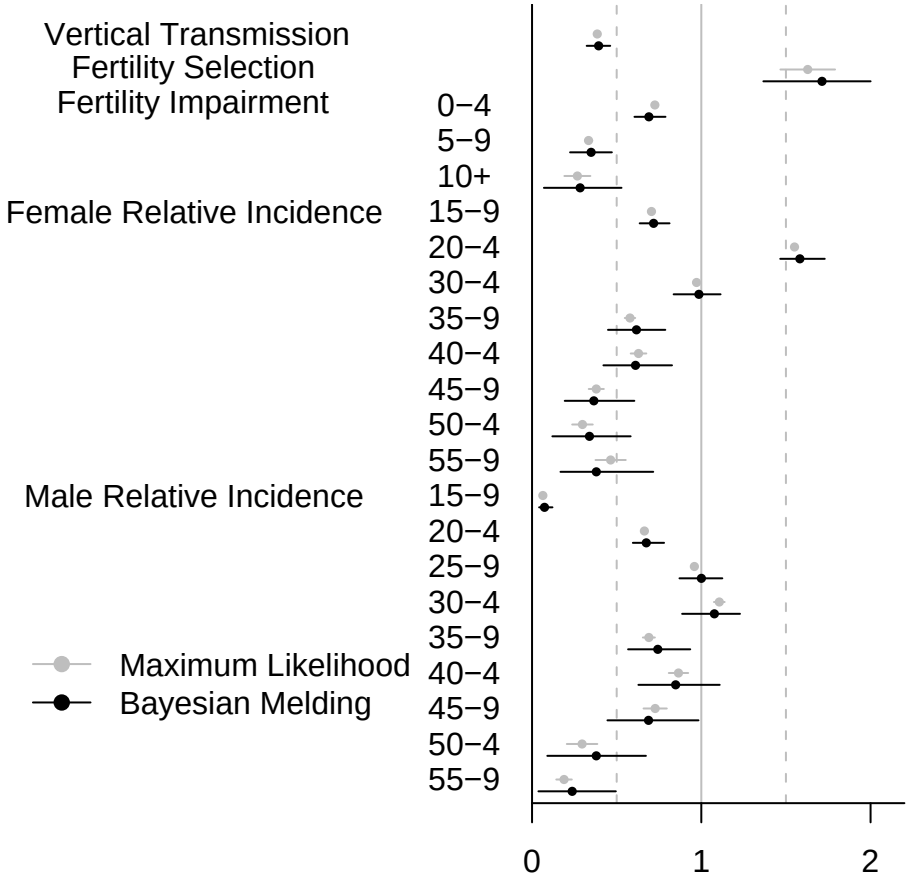


Figure 3: Posterior Samples for the Vertical Transmission, Fertility Selection, and Fertility Impairment Parameters, with Maximum Likelihood Estimates (vertical dashed lines) and Prior Distributions (horizontal lines)

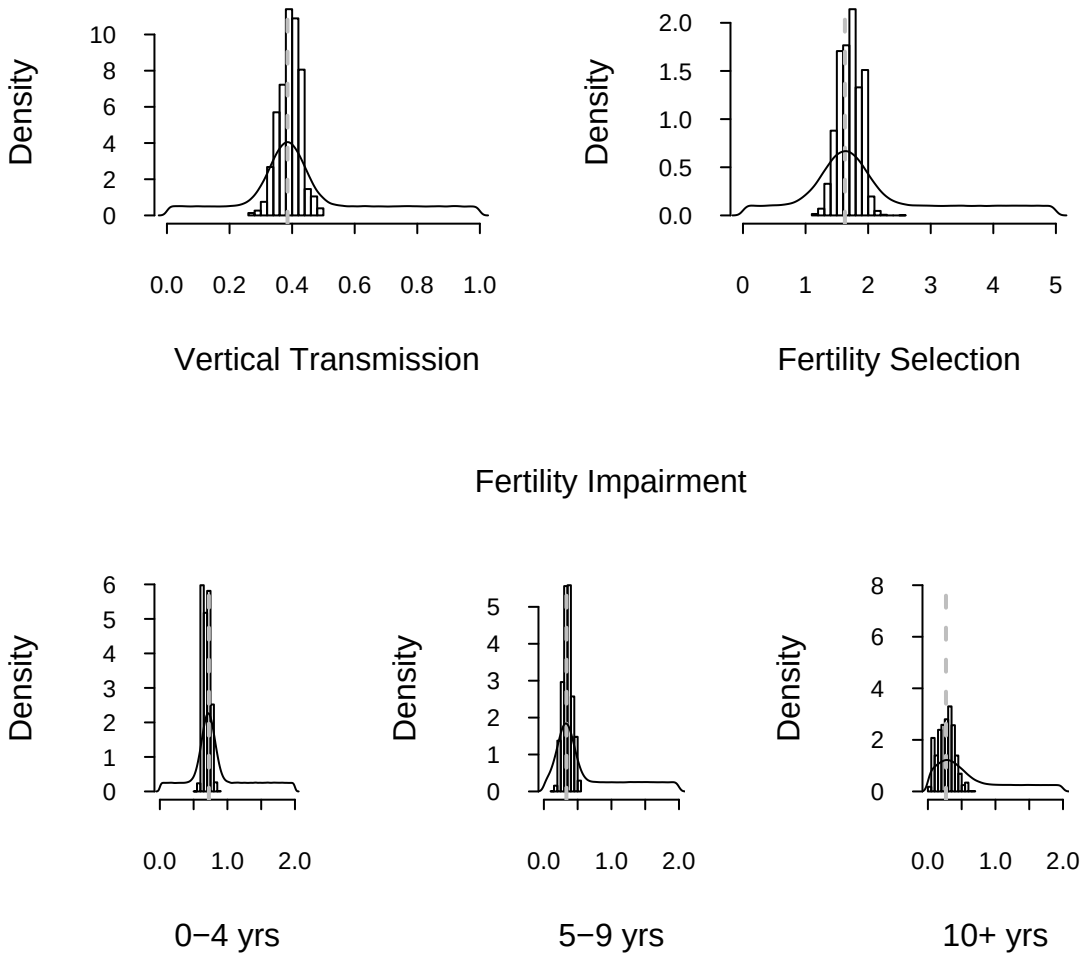


Figure 4: Posterior Samples for the Relative Incidence Ratios for Women, with Maximum Likelihood Estimates (vertical dashed lines) and Prior Distributions (horizontal lines)

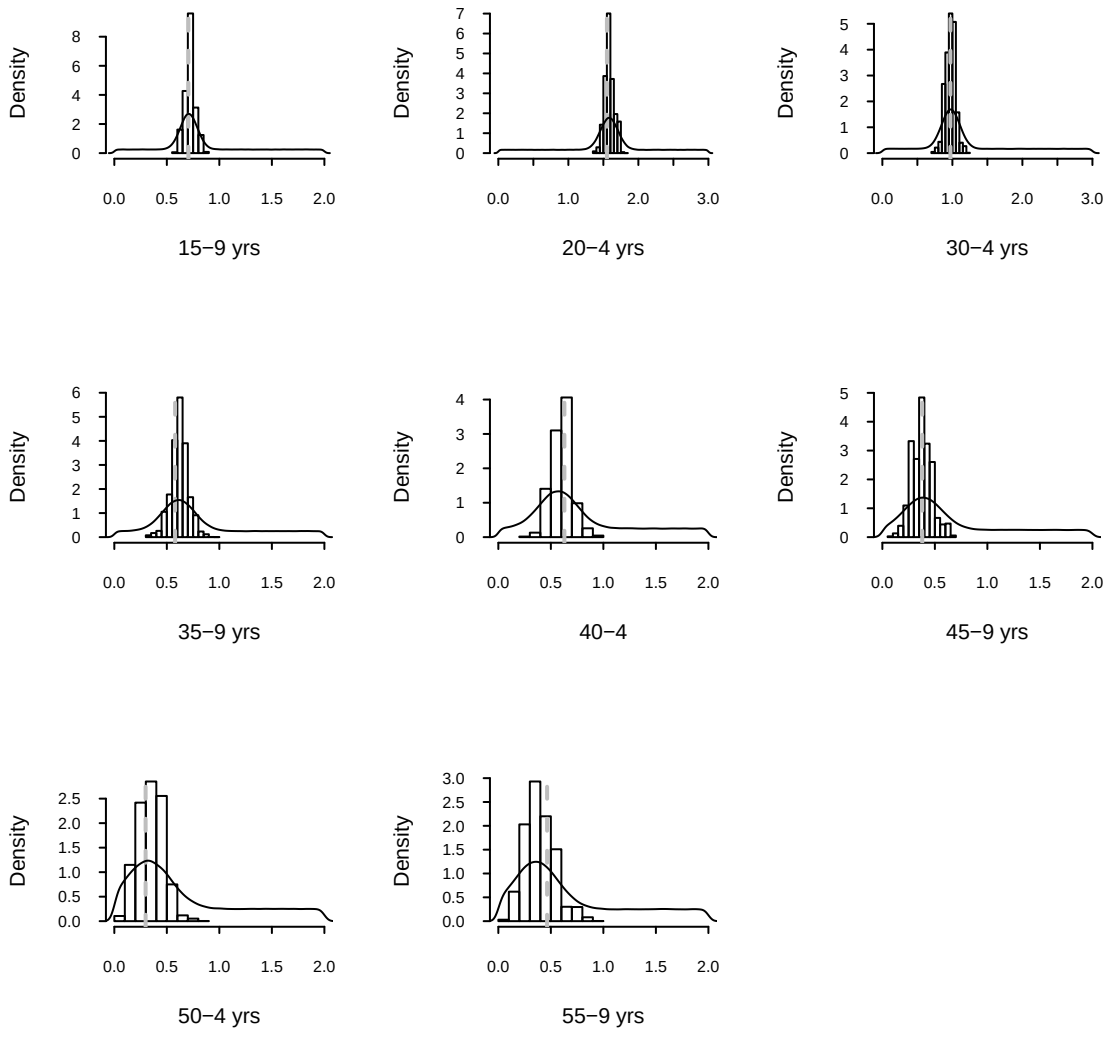


Figure 5: Posterior Samples for the Relative Incidence Ratios for Men, with Maximum Likelihood Estimates (vertical dashed lines) and Prior Distributions (horizontal lines)

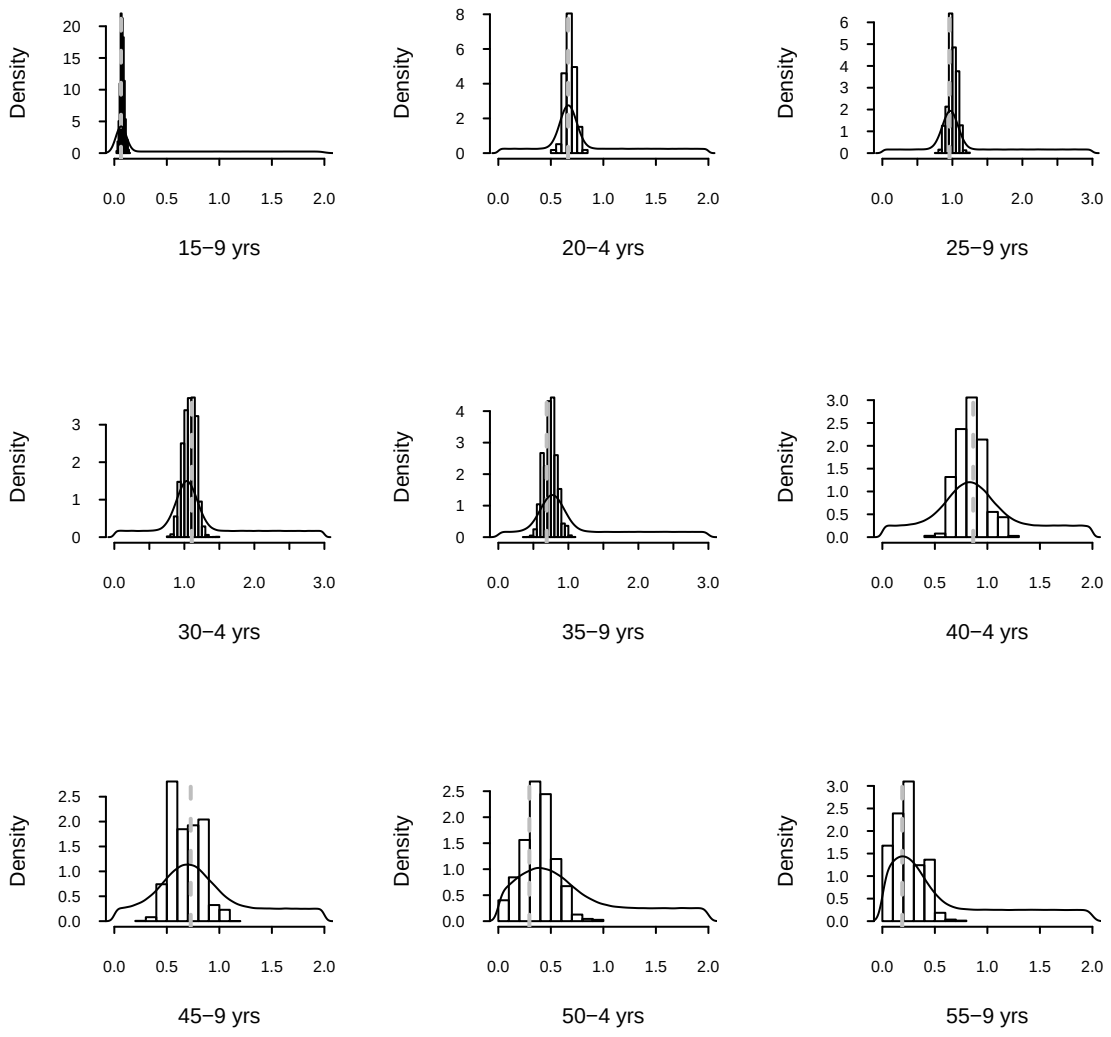


Figure 6: Posterior Distribution of Age-Specific HIV Prevalence for Women

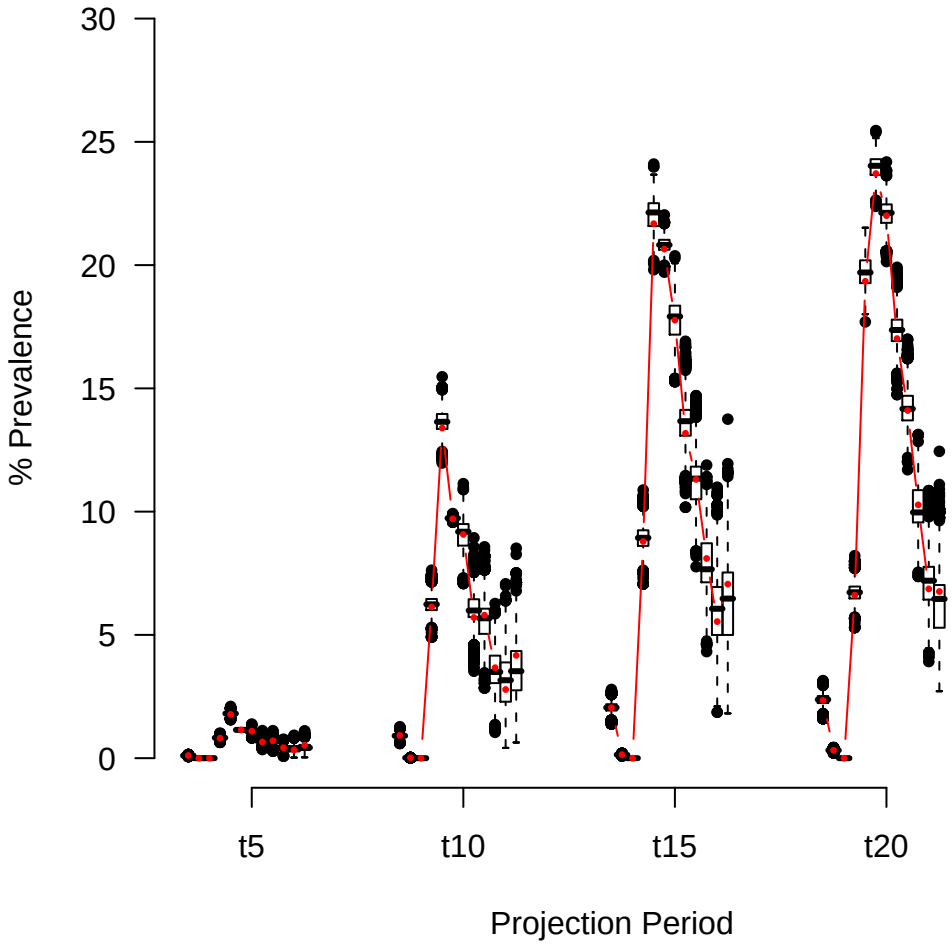
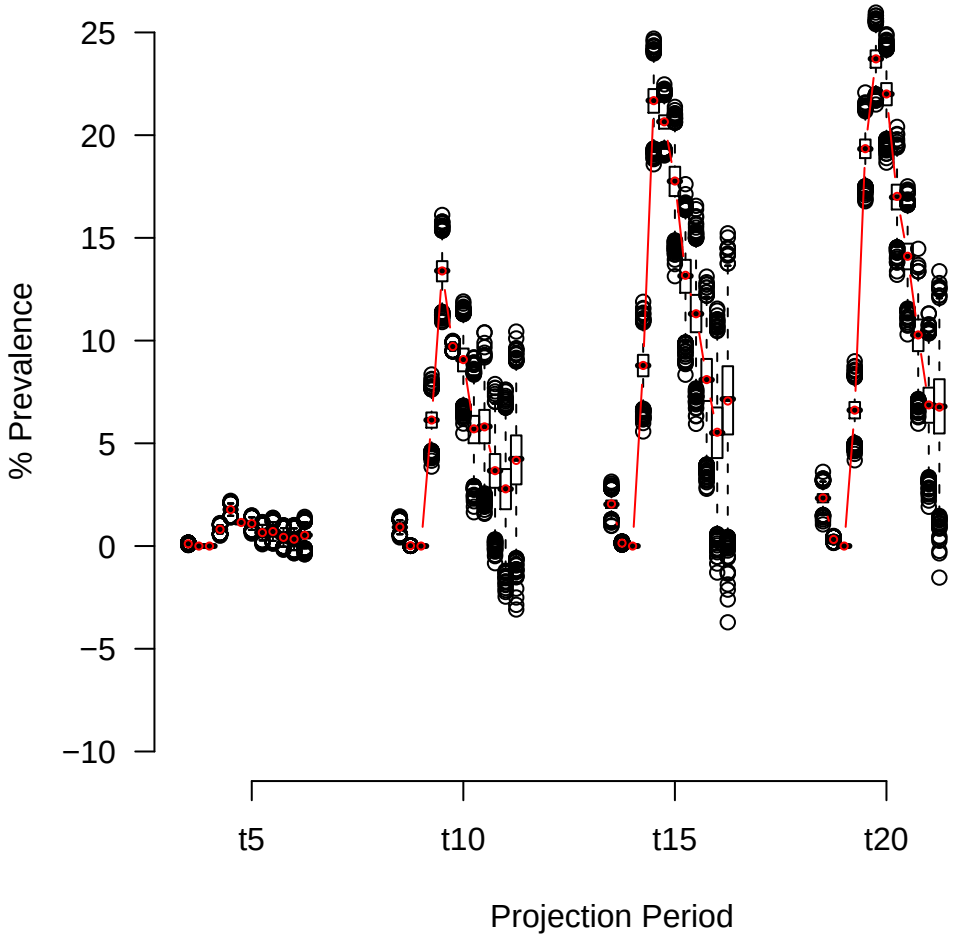


Figure 7: Age-Specific HIV Prevalence for Women Calculated from ML Estimates and 95% Confidence Intervals



TABLES

Table 1: Uniform Distributions Used in the Prior Joint Distribution for the CCMPP Parameters

Parameter	Endpoints
Vertical Transmission	0,1
Fertility Selection	0,5
Fertility Impairment	
Duration 0-4	0,2
Duration 5-9	0,2
Duration 10+	0,2
Female Relative Incidence Ratio	
15-9	0,2
20-4	0,3
25-9	0,3
30-4	0,2
35-9	0,2
40-4	0,2
45-9	0,2
50-4	0,2
55-9	0,2
Male Relative Incidence Ratio	
15-9	0,2
20-4	0,2
25-9	0,3
30-4	0,3
35-9	0,3
40-4	0,2
45-9	0,2
50-4	0,2
55-9	0,2

APPENDIX

Table 2: Maximum Likelihood Estimates for CCMPP with 95% Confidence Intervals in Parentheses.

	Heuveline	Replication	Corrected
Vertical Transmission	0.385	0.385	0.386
	(0.297, 0.478)	(0.381,0.389)	(0.381,0.39)
Fertility Selection	1.672	2.028	1.731
	(1.492, 1.865)	(1.775,2.281)	(1.569,1.893)
Fertility Impairment			
Duration 0-4	0.848	0.681	0.735
	(0.798, 0.909)	(0.672,0.689)	(0.723,0.747)
Duration 5-9	0.357	0.397	0.342
	(0.276, 0.450)	(0.384,0.409)	(0.323,0.36)
Duration 10+	0.293	0.188	0.29
	(0.078, 0.607)	(0.127,0.249)	(0.214,0.366)
Female Relative Incidence Ratio			
15-9	0.594	0.618	0.671
	(0.545, 0.650)	(0.608,0.628)	(0.663,0.679)
20-4	1.325	1.608	1.6
	(1.239, 1.412)	(1.587,1.63)	(1.581,1.619)
25-9	1.00	1.00	1.00
	–	–	–
30-4	0.752	1.036	0.998
	(0.647, 0.886)	(1.01,1.061)	(0.975,1.021)
35-9	0.635	0.571	0.576
	(0.482, 0.762)	(0.534,0.608)	(0.546,0.606)
40-4	0.551	0.679	0.609
	(0.409, 0.795)	(0.624,0.734)	(0.574,0.645)
45-9	0.356	0.435	0.421
	(0.159, 0.544)	(0.379,0.492)	(0.421,0.421)
50-4	0.295	0.331	0.312
	(0.095, 0.679)	(0.253,0.409)	(0.264,0.36)
55-9	0.246	0.553	0.445
	(0.087, 0.627)	(0.428,0.678)	(0.361,0.529)
Male Relative Incidence Ratio			
15-9	0.059	0.057	0.06
	(0.024, 0.109)	(0.056,0.058)	(0.059,0.061)
20-4	0.583	0.619	0.66
	(0.483, 0.684)	(0.61,0.628)	(0.652,0.668)
25-9	1.149	1.221	1.045
	(0.986, 1.285)	(1.193,1.249)	(1.031,1.059)
30-4	0.936	1.126	1.11
	(0.773, 1.130)	(1.073,1.178)	(1.11,1.11)
35-9	0.759	0.814	0.753
	(0.573, 0.944)	(0.757,0.872)	(0.722,0.784)
40-4	0.769	0.868	0.861
	(0.554, 1.007)	(0.787,0.949)	(0.8,0.921)
45-9	0.622	0.677	0.888
	(0.409, 0.879)	(0.595,0.759)	(0.805,0.971)
50-4	0.417	0.532	0.227
	(0.120, 0.773)	(0.401,0.663)	(0.126,0.328)
55-9	0.168	0.146	0.363
	(0.001, 0.445)	(0.105,0.187)	(0.242,0.484)