A Bayesian cohort component projection model to estimate women of reproductive age at the subnational level in data-sparse settings: Appendices

Monica Alexander* Leontine Alkema†

*University of Toronto. monica.alexander@utoronto.ca.

†University of Massachusetts, Amherst. lalkema@umass.edu.

A Other potential data sources

We use census data and WPP estimates as inputs to the model. There are other available data sources that could be used as inputs. These sources and the reasons for not including them are discussed below.

A.1 Mortality

Mortality is estimated at the subnational level based on national patterns of mortality from WPP, as well as changes in subnational population counts over time. Thus, no explicit information on subnational mortality levels is used; mortality is estimated based on likely patterns at the national level and intercensal changes in population. There are two main sources for subnational mortality data in Kenya that are not included as data inputs.

Firstly, the Demographic and Health Survey (DHS) collects information about sibling mortality histories. Adult mortality can be calculated from these data using the sibling history method, where cohorts of siblings are constructed and age-specific mortality rates are calculated based on when they died. Previous research has illustrated sibling data produces relatively reliable estimates at the national level (Masquelier, 2013). However, the DHS does not ask the location of residents of the siblings who died, thus the data cannot be used to inform differentials in subnational mortality.

A second source of information on subnational mortality comes from a question about household deaths, that was collected in the most recent census (2009). This can be used to obtain death probabilities by age. However, previous research has found that the value of ${}_{45}q_{15}$ implied by household deaths is often much lower or higher than other mortality sources (Masquelier et al. 2017). Indeed, mortality information from census household deaths is excluded from other mortality analyses due to its unreliable nature (e.g. child mortality, see UN-IGME (2017)). As such, we chose to omit this information for now. Future work will investigate this data source to see if it can be used to inform age patterns of mortality by subnational region.

A.2 Migration

There are two other potential sources of information on internal migration in Kenya that are not included as data inputs. Firstly, the census also includes a question about how many years the person has resided in their current locality of residence, referring to the district level. The question is asked in the 1999 and 2009 censuses. Based on the year of the census and the age of the respondent, as well as how many years they indicated they had lived in the current locality, the implied year and age of in-migration can be calculated. However, this method gave much lower numbers of in-migration compared to those implied by the 'location one year ago' question. As such this information was not used in the model.

Secondly, the DHS contains some information about migration.[1] For Kenya, it is possible to obtain information about the proportion of the population who moved to a particular province in the year before the survey. However, when compared to corresponding data from the census, there were large discrepancies, and trends in DHS proportions were erratic over time.

[1]Note that questions about migration in the DHS differ by country. The migration questions in the Kenya DHS are quite minimal; however for other countries there may be more useful data available.

B Full Model Specification

The full model specification is described below.

B.1 Population

B.1.1 Cohort component projection model

The underlying population by age group, year and county $\eta_{a,t,c}$ is

$$\eta_{a,t,c} = (\eta_{a-1,t-1,c} \cdot (1 - \gamma_{a-1,t-1,c})) \cdot (1 + \phi_{a-1,t-1,c}) \cdot (\varepsilon_{a-1,t-1,c}), \tag{1}$$

where $\gamma_{a,t,c}$ is the conditional probability of death in age group a, year t and county c, $\phi_{a,t,c}$ is net migration (that is, in- minus out-migration) as a proportion of population size and $\varepsilon_{a,t,c}$ is an additional age-year-county multiplier.

B.1.2 Data model

The data model is:

$$\log y_i | \eta_{a,t,c} \sim \begin{cases} N\left(\log \eta_{a[i],t[i],c[i]}, s_y^2[i]\right) & \text{if } t = 2009, \\ N\left(\log \sum_{c \in d[i]} (\eta_{a[i],t[i],c[i]}), s_y^2[i]\right) & \text{if } t < 2009, \end{cases} \tag{2}$$

where y_i is ith observed population count, s_y^2 is the sampling error based on the fact that the micro-data in IPUMS is a 10% sample. The second case of the above equation dictates that if we have observations prior to 2009, we can only relate these to $\eta_{a,t,c}$'s that have been summed to the district level.

B.1.3 National constraints

We would like to constrain the sum of the county populations to be roughly in agreement with WPP estimates, without implementing a 'hard' constraint. Ideally, we could use uncertainty around WPP estimates in this constraint, but uncertainty intervals are not published for historical WPP estimates. As such we implement the following. We constrain the sum of the county populations by

age and year to be within the interval $(\Lambda_{a,t}, \Omega_{a,t})$:

$$\Lambda_{a,t} < \sum_c \eta_{a,t,c} \leq \Omega_{a,t}, \tag{3}$$

with lower bound $\Lambda_{a,t}$ and upper bound $\Omega_{a,t}$ determined by the national estimates produced by WPP. Specifically, for the lower bound $\Lambda_{a,t}$ we assume the following prior:

$$\log \Lambda_{a,t} \sim N(\log 0.9 WPP_{a,t}, 0.1^2) T(, \log WPP_{a,t}). \tag{4}$$

This prior dictates that the prior probability of $\sum_c \eta_{a,t,c} < 0.9 WPP_{a,t}$ is 50%. The standard deviation of 0.1 on the log-scale captures the uncertainty associated with the lower bound. We assign a WPP-informed prior to upper bound $\Omega_{a,t}$ in a similar manner:

$$\log \Omega_{a,t} \sim N(\log 1.1 WPP_{a,t}, 0.1^2) T(\log WPP_{a,t},). \tag{5}$$

Note that in this set-up, we do not use WPP estimates as 'data' to directly inform the sum of the county estimates as in other work (ref to world pop work where wpp feeds into the likelihood function directly). Instead, we use the WPP estimates to exclude combinations of that are extreme as compared to the WPP estimates.

B.1.4 Priors on first year and age group

The cohort component projection framework requires priors to be placed on populations in the first year and age group. We use the following priors:

$$\log \eta_{1,t,c} \sim N(\log WPP_{1,t} + \log \text{prop}_{1,t,c}, 0.01^2), \tag{6}$$

$$\log \eta_{a,1,c} \sim N(\log WPP_{a,1} + \log \text{prop}_{a,1,c}, 0.01^2), \tag{7}$$

where $WPP_{a,t}$ is the national-level population count from WPP in the relevant age group and year, and $\text{prop}_{a,t,c}$ is the proportion of the total population in the relevant age, year and county, which was calculated based on interpolating census year proportions and assuming the proportion of a district's population in each county was constant at a level equal to 2009.

B.2 Mortality

The model for mortality is as

$$\text{logit}\gamma_{a,t,c} = Y_{a,0} + \beta_{t,c,1} \cdot Y_{a,1} + \beta_{t,c,2} \cdot Y_{a,2}, \tag{8}$$

where $Y_{a,0}$ is the mean age-specific logit mortality schedule of the national mortality curves and $Y_{,1}$ and $Y_{,2}$ are the first two principal components derived from national-level mortality schedules. Modeling on the logit scale ensures the death probabilities are between zero and one.

The county-specific coefficients $\beta_{t,c,k}$ are modeled as fluctuations around a national mean:

$$\beta_{t,c,k} = B^{nat}_{t,k} + \delta_{t,c,k}, \tag{9}$$

$$\delta_{t,c,k}|\delta_{t-1,c,k}, \sigma^2_\delta \sim N(\delta_{t-1,c,k}, \sigma^2_\delta), \tag{10}$$

where $B^{nat}_{a,t,k}$ are the national coefficient on principal components, derived from WPP data. The county-specific fluctuations are modeled as a random walk.

B.3 Migration

B.3.1 Process model

The process model for the net-migration is:

$$\phi_{a,t,c} = \frac{\psi^{in}_{a,t,c} - \psi^{out}_{a,t,c}}{\eta_{a-1,t-1,c}}, \tag{11}$$

$$\psi^{in}_{a,t,c} = \Psi^{in}_{t,c} \cdot \Pi^{in}_{a,c}, \tag{12}$$

$$\psi^{out}_{a,t,c} = \Psi^{out}_{t,c} \cdot \Pi^{out}_{a,c}, \tag{13}$$

where $\Psi^{in}_{t,c}$ and $\Psi^{out}_{t,c}$ are the total number of in- and out-migrants, respectively, and $\Pi^{in}_{a,c}$ and $\Pi^{out}_{a,c}$ are the relevant age distributions. We model the total counts as a second order random walk to

impose a certain level of smoothness in the counts over time:

$$\begin{aligned}
\Psi^{in}_{1,c} &\sim U(0, y_c), && (14)\\
\log \Psi^{in}_{2,c} | \Psi^{in}_{1,c}, \sigma^2_{in} &\sim N(\log \Psi^{in}_{1,c}, \sigma^2_{in}), && (15)\\
\log \Psi_{t,c^{in}} | \Psi^{in}_{(t-2):(t-1),c}, \sigma^2_{in} &\sim N(2 \log \Psi^{in}_{t-1,c} - \log \Psi^{in}_{t-2,c}, \sigma^2_{in}), && (16)\\
\Psi^{out}_{1,c} &\sim U(0, y_c), && (17)\\
\log \Psi^{out}_{2,c} | \Psi^{out}_{1,c}, \sigma^2_{out} &\sim N(\log \Psi^{out}_{1,c}, \sigma^2_{out}), && (18)\\
\log \Psi^{out}_{t,c} | \Psi^{out}_{(t-2):(t-1),c}, \sigma^2_{out} &\sim N(2 \log \Psi^{out}_{t-1,c} - \log \Psi^{out}_{t-2,c}, \sigma^2_{out}). && (19)
\end{aligned}$$

where y_c refers to the observed total population for county c based on the census in the first observation period.

We place Uniform priors on the non-normalized age distributions of in- and out-migration, with equal prior probability on each age group:

$$\begin{aligned}
\Pi^{in*}_{a,c} &\sim \text{Uniform}(0, 1), && (20)\\
\Pi^{out*}_{a,c} &\sim \text{Uniform}(0, 1). && (21)
\end{aligned}$$

We then normalize the age distributions as

$$\begin{aligned}
\Pi^{in}_{a,c} &= \frac{\Pi^{in*}_{a,c}}{\sum_a \Pi^{in*}_{a,c}}, && (22)\\
\Pi^{out}_{a,c} &= \frac{\Pi^{out*}_{a,c}}{\sum_a \Pi^{out*}_{a,c}}. && (23)
\end{aligned}$$

B.3.2 Data model

We relate the observed age-specific in- and out-migration counts in the censuses, denoted M_i^{in} and M_i^{out}, respectively, to the underlying true counts $\psi_{a,t,c}^{in}$ and $\psi_{a,t,c}^{out}$ through the following data model:

$$\log M_i^{in} | \psi_{a,t,c}^{in} \sim \begin{cases} N\left(\log \psi_{a[i],t[i],c[i]}^{in}, s_{in}^2[i]\right) & \text{if } t[i] = 2009, \\ N\left(\log \sum_{c \in d[i]} (\psi_{a[i],t[i],c[i]}^{in}), s_{in}^2[i]\right) & \text{if } t[i] < 2009, \end{cases} \tag{24}$$

$$\log M_i^{out} | \psi_{a,t,c}^{out} \sim \begin{cases} N\left(\log \psi_{a[i],t[i],c[i]}^{out}, s_{out}^2[i]\right) & \text{if } t[i] = 2009, \\ N\left(\log \sum_{c \in d[i]} (\psi_{a[i],t[i],c[i]}^{out}), s_{out}^2[i]\right) & \text{if } t[i] < 2009. \end{cases} \tag{25}$$

B.3.3 Constraint

We implement the following constraint:

$$\sum_c -0.1\eta_{a,t,c} < \sum_c \psi_{a,t,c}^{in} - \sum_c \psi_{a,t,c}^{out} \leq \sum_c 0.1\eta_{a,t,c}. \tag{26}$$

The constraint states that the difference between the sum of all in- and out-migration flows across all counties cannot be more than $\pm 10\%$ of the total estimated national population for that particular age group and year.

B.4 Age-time multiplier

We model multipliers on the log scale, and to ensure identifiability we assume the mean of the sum of the log multipliers is zero. This constraint is implemented through the re-parameterization:

$$\log \boldsymbol{\varepsilon}_{1:A,t,c} = \mathbf{D}(\mathbf{D}\mathbf{D}')^{-1} \boldsymbol{\zeta}_{1:(A-1),t,c}, \tag{27}$$

$$\boldsymbol{\zeta}_{a,t,c} \sim N(0, \sigma_\zeta^2), \tag{28}$$

where $\mathbf{D}$ is first-order difference matrix (with $D_{i,i} = -1$, $D_{i,i+1} = 1$, and $D_{i,j} = 0$ otherwise) such that $\zeta_{a,t,c} = \log \varepsilon_{a,t,c} - \log \boldsymbol{\varepsilon}_{a-1,t,c}$.

B.5 Priors on variance parameters

All variance parameters that are estimated ($\sigma^2_\alpha, \sigma^2_\delta, \sigma^2_\Psi, \sigma^2_{in}$ and σ^2_{out}) have half-Normal standard priors placed on them, i.e.

$$\sigma \sim N^+(0, 1).$$

C Age patterns in migration data

In the Bayesian cohort component model, specifically in the migration process model, we assume the age distribution of in- and out-migrants by county is constant over time (see Equations 13 and 14). This is a somewhat strong assumption and was made to ensure identifiability of all parameters in the model in cases where we do not have very much data. While the assumption is relatively strong, it was motivated by age patterns observed in census data. Figures 1 and 2 show the proportion of all in- and out-migrants by age group for each year and district, and illustrate that the age patterns remain remarkably constant over time. For reference, the broad areas covered by the IPUMS districts are listed in Table 1. Note that in addition to age distributions being constant over time, Figures 1 and 2 show they are also quite similar across both in and out flows, so it would be possible to simplify the model even further to just have one age distribution for each county that is constant across time and the same for both flow directions.

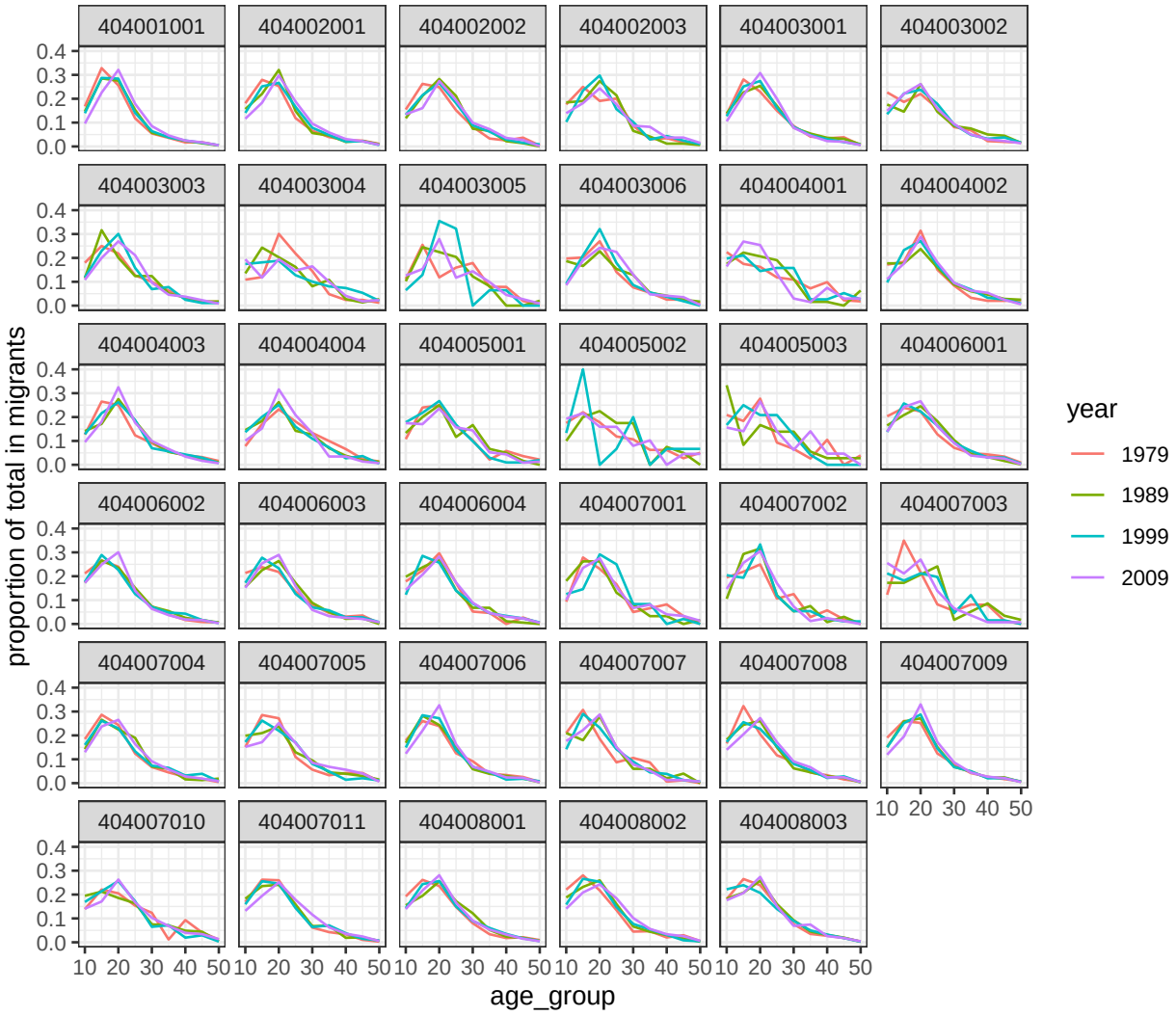


Figure 1: Observed age patterns of in-migration from Kenyan censuses, 1979-2009.

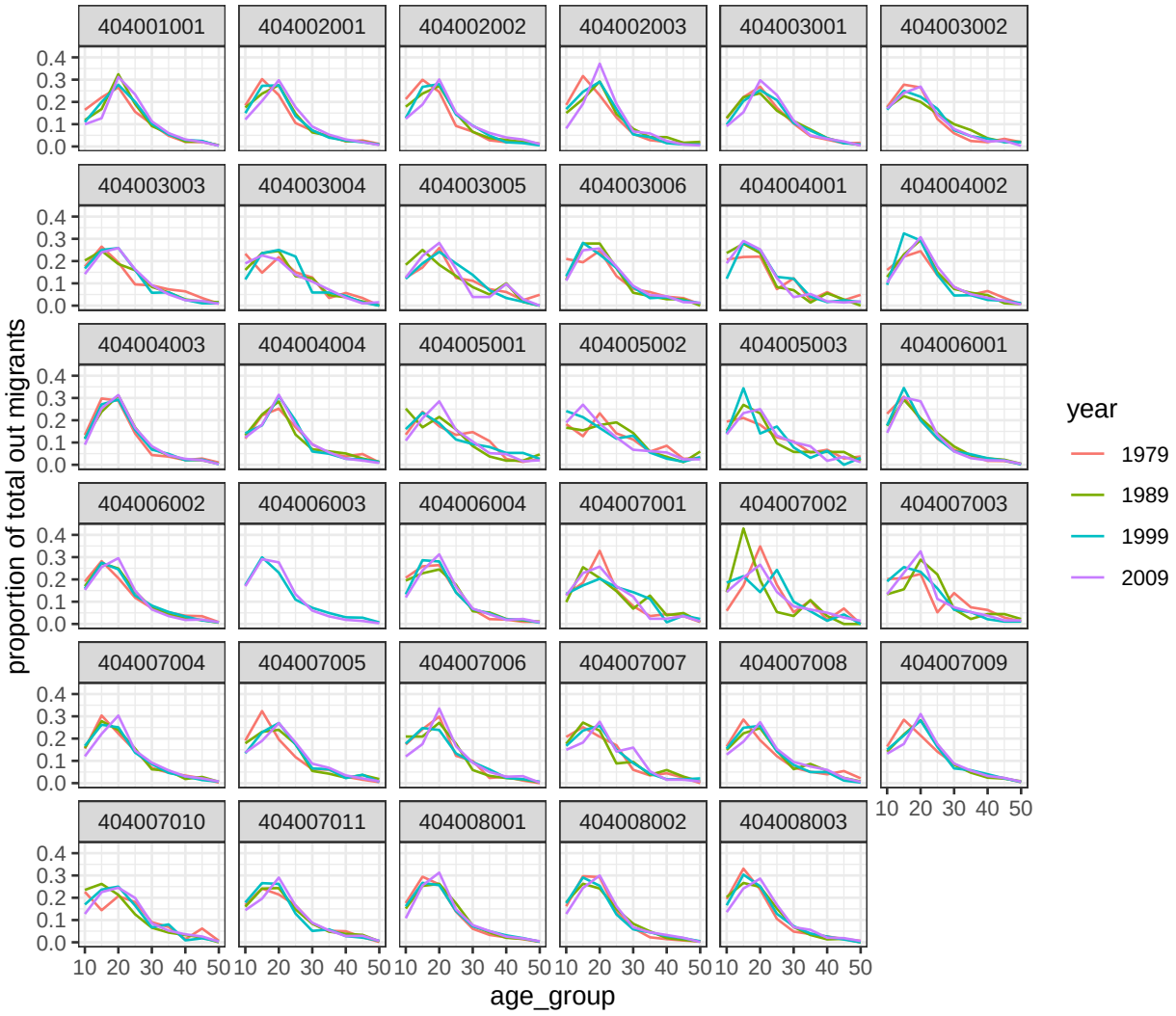


Figure 2: Observed age patterns of out-migration from Kenyan censuses, 1979-2009.

District	Areas
404001001	Nairobi East, Nairobi North, Nairobi West, Westlands
404002001	Gatanga, Gatundu, Githunguri, Kiambu (Kiambaa), Kikuyu, Lari, Muranga, Nyandarua, Ruiru, Thika, Maragua
404004001	Chalbi, Laisamis, Marsabit, Moyale
404004002	Garba Tulla, Igembe, Imenti, Isiolo, Maara, Meru, Tharaka, Tigania, Meru
404004003	Embu, Kangundo, Kibwezi, Machakos, Makueni, Mbeere, Mbooni, Mwala, Nzaui, Yatta
404004004	Kitui North, Kitui South (Mutomo), Kyuso, Mwingi
404005001	Fafi, Garissa, Ijara, Lagdera
404005002	Wajir East, Wajir North, Wajir South, Wajir West
404005003	Mandera Central, Mandera East, Mandera West
404006001	Bondo, Rarieda, Siaya
404006002	Kisumu East, Kisumu West, Nyando
404006003	Homa Bay, Kuria East, Kuria West, Migori, Rachuonyo, Rongo, Suba
404002002	Nyeri North, Nyeri South
404006004	Borabu, Gucha, Gucha South, Kisii Central, Kisii South, Manga, Masaba, Nyamira, North Kisii
404007001	Turkana Central, Turkana North, Turkana South
404007002	Pokot Central, Pokot North, West Pokot
404007003	Samburu Central, Samburu East, Samburu North
404007004	Kwanza, Trans Nzoia East, Trans Nzoia West
404007005	Baringo, Baringo North, East Pokot, Koibatek, Laikipia East, Laikipia North, Laikipia West
404007006	Eldoret East, Eldoret West, Wareng, Uasin Gishu
404007007	Keiyo, Marakwet, Elgeyo Markwet
404007008	Nandi Central, Nandi East, Nandi North, Nandi South, Tinderet
404007009	Kaijiado Central, Kaijiado North, Loitoktok, Molo, Naivasha, Nakuru, Nakuru North, Kajiado
404002003	Kirinyaga
404007010	Narok North, Narok South, Trans Mara
404007011	Bomet, Buret, Kericho, Kipkelion, Sotik
404008001	Butere, Emuhaya, Hamisi, Kakamega, Lugari, Mumias, Vihiga, Butere/Mumias
404008002	Bungoma East, Bungoma North, Bungoma South, Bungoma West, Mt. Elgon
404008003	Bunyala, Busia, Samia, Teso North, Teso South
404888001	Waterbodies
404003001	Kilindini, Kilindini, Mombasa
404003002	Kinango, Kwale, Msambweni
404003003	Kaloleni, Kilifi, Malindi
404003004	Tana Delta, Tana River
404003005	Lamu
404003006	Taita, Taveta, Taita Taveta

Table 1: IPUMS district codes and areas covered

D Mortality Principal Components

We included a mean mortality schedule and two principal components in our model for mortality. The mean schedule and first three principal components are showing in Figure 3. A broad demographic interpretation is the first component shows overall mortality change and the second principal component is the main component representing HIV/AIDS. This is slightly different to previous work (Sharrow et al. 2014) because we perform the SVD on the demeaned mortality schedules. Note that the first two principal components account for over 91% of variation across all schedules. We experimented with including the third principal component but the resulting coefficient estimates were generally very close to zero so decided against including it.

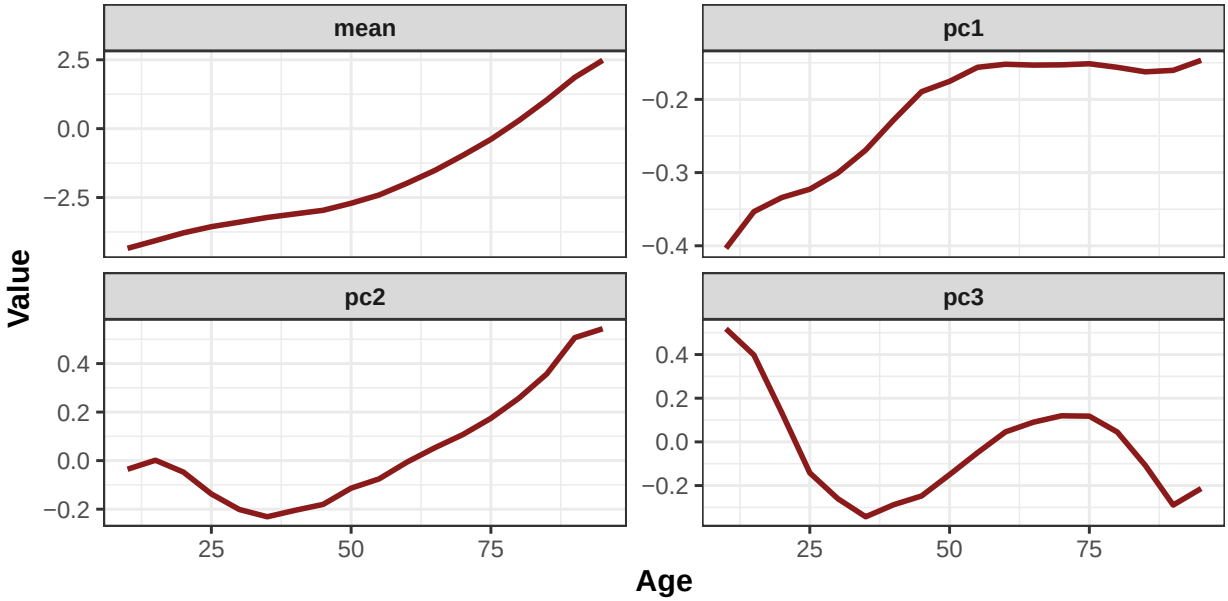


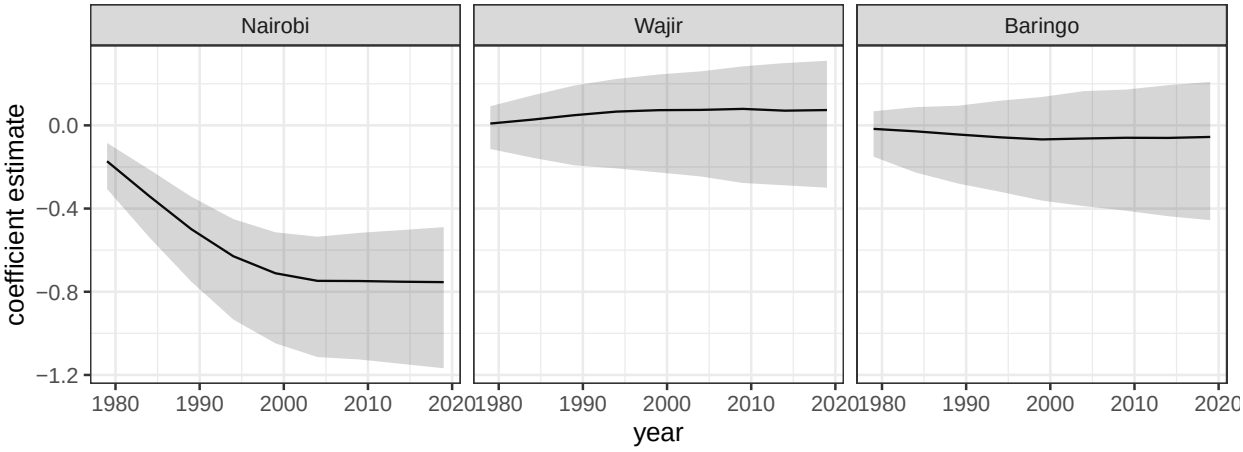
Figure 3: Mean mortality schedule and first three principal components derived from WPP life table estimates for Sub-Saharan African countries.

E Additional results

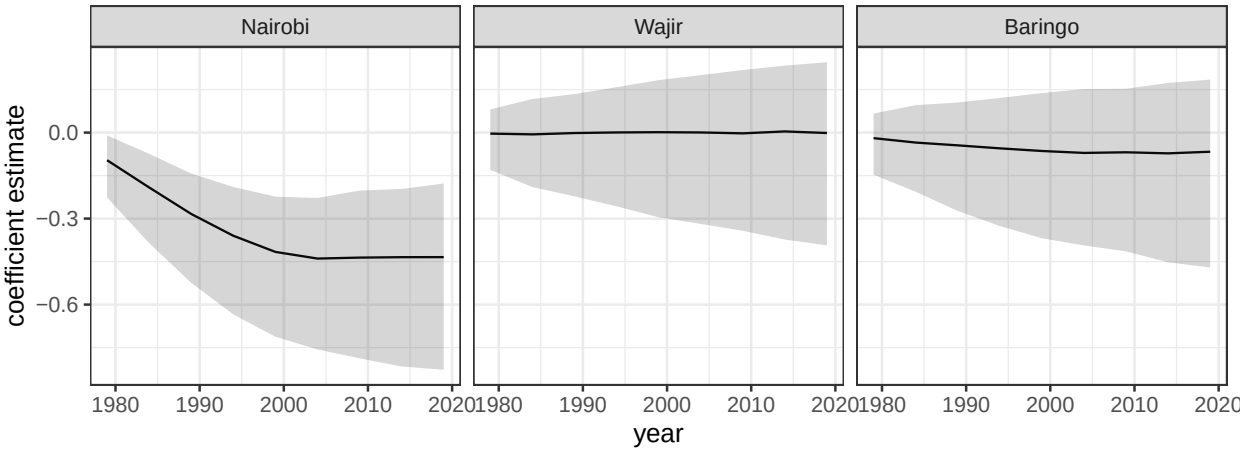
In this section we highlight several other components that are estimated within the model; specifically the coefficients on the first and second principal components. Results are illustrated on three example counties: Nairobi, Wajir and Baringo. Additionally, we show estimates for the age-time multiplier for all counties.

Figure 4 shows estimates over time of the coefficient of the first and second principal component within the mortality model (i.e. $\beta_{tc,1}$ and $\beta_{tc,2}$). Broadly, the first principal component relates to overall mortality improvement, and the second relates to the effect of the HIV/AIDS epidemic. Coefficients on the first component suggest mortality improvement is relatively slow in Nairobi, but close to average in the second two counties. Based on patterns on the second principal component, there is evidence to suggest that the effect of HIV/AIDS epidemic was relatively high in Nairobi (Figure 4).

Figure 5 shows the estimated age-time specific multiplier for all counties. As can be seen, the estimates on the log scale are very close to zero for the majority of age groups, years and counties.



(a) Estimates for first component (region deviations from national mortality trends)



(b) Estimates for second component (region deviations from national HIV/AIDS mortality)

Figure 4: County-specific deviations from national-level mortality improvements (first component) and HIV/AIDS mortality (second component) for three counties.

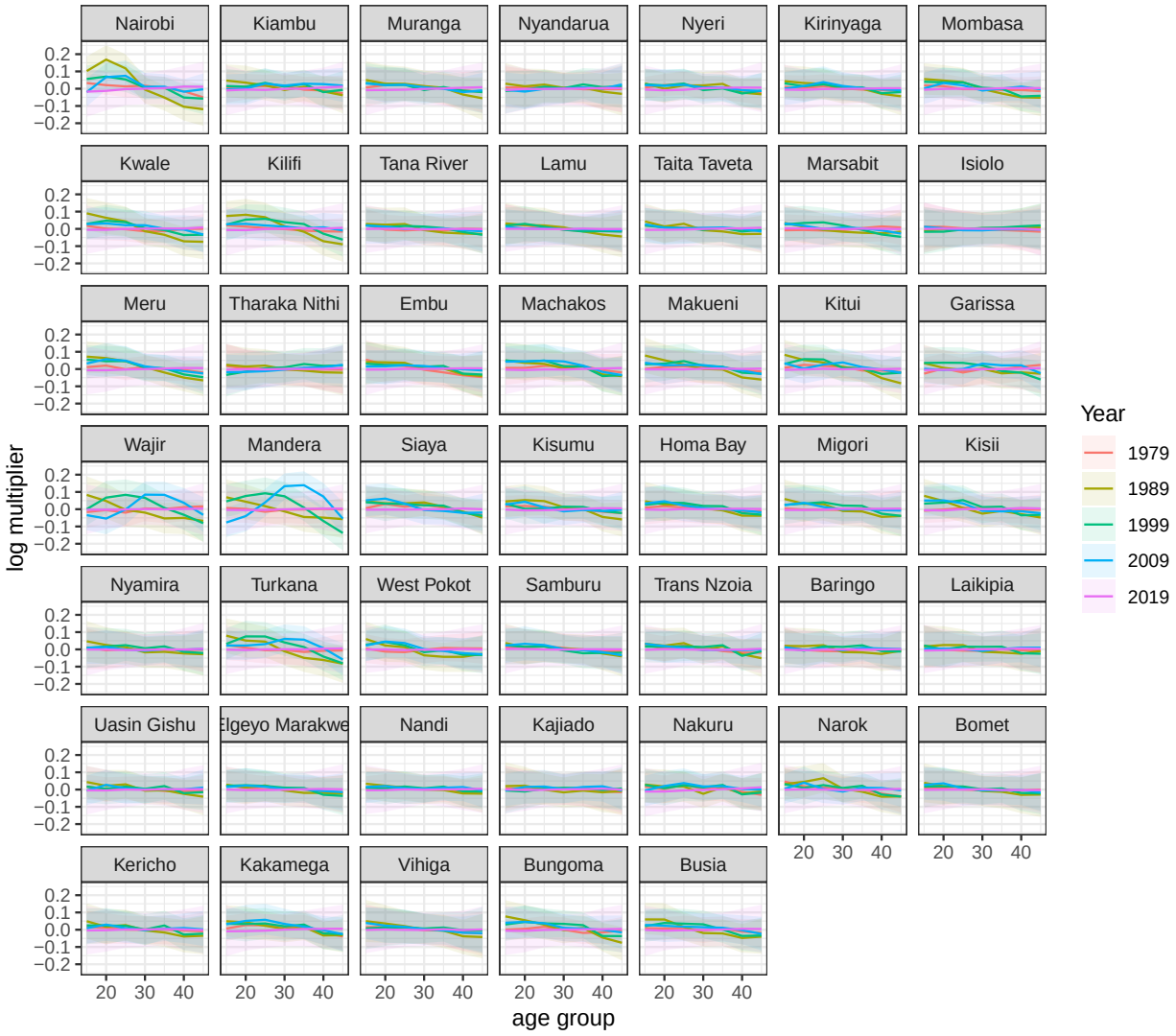


Figure 5: Age-time specific multipliers for all counties.

F PIT histogram

A Probability Integral Transform (PIT) histogram is a tool for evaluating the similarity between model projections and left out observations. The predictive distributions of the projections are compared with the actual observations (Angus 1994).

For each observation j for 2019 (i.e. each population count by age group and county) we have observation y_j from the 2019 census, and sample $\hat{\eta}_j^{(S)}$ from the corresponding posterior distribution (with a total of S samples). The PIT for observation j was calculated as

$$PIT_j = \frac{\sum_{s=1}^{S} \hat{\eta}_j^{(s)} \leq y_j}{S}. \tag{29}$$

If the predictive distribution is well calibrated, the result should be a uniform distribution of PIT values. Figure 6 shows the PIT histogram for 2019. The relatively high density in the middle of the distribution suggests the model is somewhat over-dispersed, and the low density towards 1 suggests the upper bound of population projections is in general too conservative.

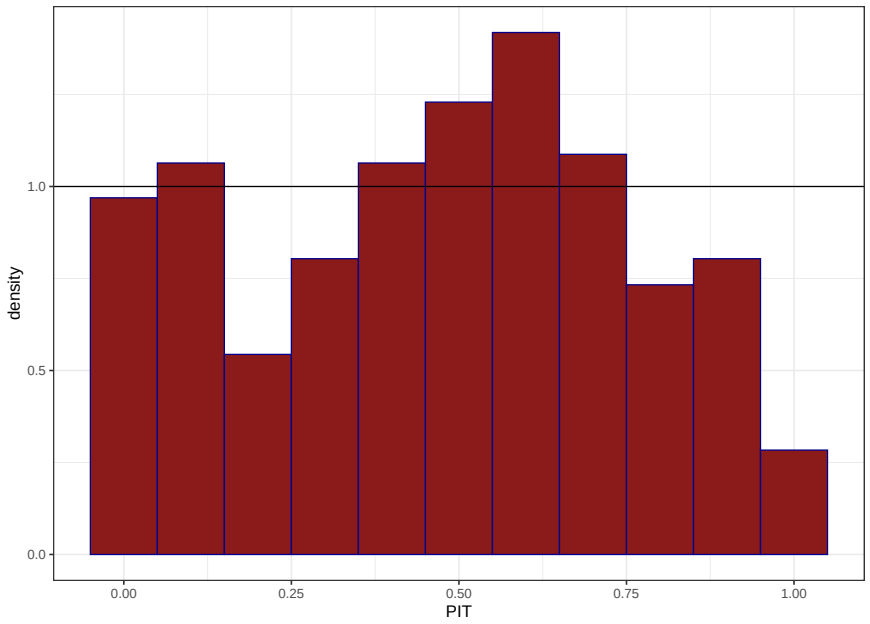


Figure 6: PIT histogram comparing projected 2019 population counts with observed 2019 census counts.