A Bayesian hierarchical model to estimate subnational populations of women of reproductive age

Monica Alexander[*]
Leontine Alkema[†]

Paper presented at PAA 2018

Abstract

Accurate estimates of subnational populations are important for policy formulation and monitoring population health indicators. In particular, estimates of the number of women of reproductive age affect measures of maternal mortality, contraceptive prevalence and fertility. However, in many developing countries, data on population counts are limited and are of poor quality, and so levels are unclear. We present a Bayesian hierarchical model to estimate female populations at the subnational level. The model builds on a cohort component projection framework, incorporates data on population counts and migration, and uses characteristic mortality schedules to obtain population estimates and uncertainty levels. The model is applied to estimate and project populations by county in Kenya for 1979-2020.

1 Introduction

Reliable estimates of demographic and health indicators at the subnational level are essential for monitoring trends and inequalities over time. As progress towards targets such as the Sustainable Development Goals (SDGs) is tracked, there has been increasing recognition of the substantial differences that can occur across regions within a country (World Health Organization (WHO) (2016b); Lim et al. (2016); He et al. (2017)). It is important to measure and monitor trends at the subnational level to fully understand a country's progress and likely future trajectories.

To effectively measure health indicators of interest, we need to be able to accurately estimate the size of the population at risk. Differences in estimates of the denominator can have a large effect on the resulting estimates of key indicators. For example, in 2017 the United Nations Inter-agency Group for Child Mortality Estimation (UN-IGME) and the Institute for Health Metrics and Evaluation (IHME) both published estimates of under-five child mortality in countries worldwide (GBD 2016 Mortality Collaborators (IHME) (2017); UN-IGME (2017)). However, estimates for 2016 differed markedly, with IHME's estimate being 642,000 deaths lower than the UN-IGME estimate. The main reason for the discrepancy was the different sets of estimates of live births: IHME assumed there were 128.8 million live births in 2016, which was 12.2 million lower than the 141 million used by UN-IGME. Thus, it is important to accurately measure the population at risk over time.

A particular population of interest is women of reproductive age (WRA), i.e. those aged 15-49. This subgroup forms the population at risk for many important health indicators such as fertility rates, maternal mortality, and measures of contraceptive prevalence. We need to be able to accurately estimate the size of the population at risk in order to effectively measure these indicators. However, the data available on the number of WRA at the subnational level vary substantially by country. Often data availability and quality is the worst in countries where outcomes are also relatively poor. For example, many developing countries may only have one or two historical censuses available. Although simple population interpolation and projection is often possible using the available data, these methods do not account for changing mortality or migration patterns, and do not give any indication of uncertainty around the estimates or projections. As such, we need to employ statistical models to come up with robust population estimates and uncertainty levels.

[*]University of California, Berkeley. monicaalexander@berkeley.edu.
[†]University of Massachusetts, Amherst. lalkema@umass.edu.

In this paper, we present a Bayesian hierarchical model to estimate subnational populations of WRA. The model embeds a cohort component projection setup in a Bayesian framework, allowing uncertainty in data and population processes to be taken into account. The model uses available data on population and migration counts from censuses, as well as national-level information on mortality and population trends. As such, the methodology is applicable across a wide range of countries. Subnational estimates are calibrated to produce results that are consistent with those produced by the UN as part of the World Population Prospects (UNPD (2017a)). Results from the model can input to demographic and health indicators at the subnational level but also to understand drivers of population change and how these may in turn affect trends in indicators of interest.

The remainder of this paper is structured as follows. The next section gives a brief overview of existing methods of subnational population estimation. We then describe the main data sources used, and then give a detailed description of the proposed methodology. The performance of the model is then illustrated through the estimation and projection of populations across districts in Kenya. Results and future work are then discussed.

2 Existing methods of subnational population estimation

Methods to estimate population at the subnational level are similar to estimation methods at the national level. However, there are several notable challenges of subnational population estimation that do not exist at a country level. Firstly, migration flows are more important at the subnational level. While migration flows are often assumed to be negligible at the national level, they are usually larger as a proportion of total population size at the regional level. In addition, migration flows at the subnational level are also often more difficult to estimate. Any particular region could have net in- or out-migration, and flows to and from different regions can differ markedly in magnitude. Secondly, when estimating subnational populations, it is important to ensure the sum of all regions agrees with national estimates produced elsewhere. In practice, this usually involves a process of calibration against a known national population, adjusting shares by region in potentially an ad-hoc way, so that they match the total. Lastly, data quality and availability is often poorer at the subnational level. Populations at the regional level are smaller and data are often more volatile, and data on key indicators of mortality and internal migration is often lacking or unreliable.

2.1 Traditional methods

Perhaps the simplest and least data-intensive methods of subnational population estimation involve interpolation and extrapolation of regional shares of the total population (Swanson and Tayman (2012)). Given two (or more) censuses, one can calculate the relevant shares of the population by age, sex and region and see how they have changed over time. Intercensal estimations of populations assume constant increase (or decrease) over time. Projection of populations into the future can then be made based on assumptions of constant levels or trends in shares. For example, the U.S. Census Bureau produce subnational population estimates for the majority of countries worldwide (U.S. Census Bureau (2017)). The methods used to produce such estimates involve making assumptions such as constant or logistic growth, and iteratively calculating population proportions by age, sex and region such that they match the country's total populations (Leddy (2017)).

The most commonly used methods of population estimation and projection are cohort component methods. These center on the demographic accounting identity, which states that the population size (P) at time t is equal to the population size at $t-1$, plus births (B) and in-migrants (I), minus deaths (D) and out-migrants (O) (Wachter (2014)):

$$P_t = P_{t-1} + B_{t-1} + I_{t-1} - D_{t-1} - O_{t-1} \tag{1}$$

The above equation is for a total population, but the same accounting equation holds for each age group separately (where births only affect the first age group). The cohort component method of population projection (Leslie (1945)) takes a baseline population with a certain age structure and survives it forward based on age-specific mortality, fertility and migration rates. Cohort component methods are more data-intensive than extrapolation methods, which is particularly an issue at the subnational level. For developing countries in particular, where well-functioning vital registration systems do not exist, sufficient data on mortality, fertility and migration is often lacking.

Other methods of subnational estimation involve building regression models which relate other variables of interest to changes in population over time. For example, one could regress the ratio of census populations (area of interest / total population) against the ratio of some other indicator e.g. births, deaths, voters, school enrollments (see Swanson and Tayman (2012) for a detailed review). However, given the lack of data available in many developing countries – on population counts, let alone other indicators of growth – these methods have limited use in our context.

These traditional methods of population estimation are deterministic and do not account for random variation in demographic processes and possible measurement errors that may exist in the data. In practice, the population data that are available in developing countries are often sparse and may suffer from various types of errors. When estimating and projecting population sizes through time, it is particularly important in developing country contexts to give some indication of the level of uncertainty around those estimates, based on stochastic error, measurement error and uncertainties in the underlying modeling process.

2.2 Bayesian methods

The use of Bayesian methods in demography has become increasingly common, as it provides a useful framework to incorporate different data sources in the same model, account for various types of uncertainty, and allow for information exchange across time and space (Bijak and Bryant (2016)). Bayesian methods have been used to model and forecast national populations (Raftery et al. (2012); UNPD (2017a)), fertility (Alkema et al. (2011)), mortality (Alexander and Alkema (2018); Alkema and New (2014); Girosi and King (2008)) and migration (Bijak (2008)). Particularly relevant to this work is the methodology proposed by Wheldon et al. (2013) for the reconstruction of past populations. The model embeds the demographic accounting equation within a Bayesian hierarchical framework, using information from available censuses and surveys to reconstruct historical populations. The authors show the method works well to estimate populations and quantify uncertainty in a wide range of countries with varying data availability (Wheldon et al. (2016)).

In the field of subnational estimation, Bayesian methods have also been used in many different contexts. For subnational mortality estimation, many researchers have used Bayesian hierarchical frameworks to share information about mortality trends across space and time, in contexts where the available data are both reliable (Congdon, Shouls, and Curtis (1997); Alexander, Zagheni, and Barbieri (2017)) and sparse (Schmertmann and Gonzaga (2018)). For subnational fertility estimation, Sevcikova, Raftery, and Gerland (2017) propose a Bayesian model that produces estimates and projections of subnational total fertility rates (TFRs) that are consistent with national estimates of TFR produced by the UN. Building from the local level up, Schmertmann et al. (2013) propose a method which uses empirical Bayesian methods to smooth volatile fertility data at the regional level, before modeling using a Brass relational model variant. In terms of population estimation at the subnational level, Bryant and Graham (2013) proposed a Bayesian hierarchical model to estimate subnational populations in New Zealand. Similarly to the Wheldon et al. methodology, this model builds a framework around the demographic accounting equation, allowing information from many different sources The focus of the Bryant and Graham paper is how to reconcile and incorporate information about the population from sources such as censuses, and school and voting enrollments.

There is an increasing amount of work using geo-located data and satellite imagery to estimate population sizes and flows in developing countries (Tatem et al. (2013)). Led by the WorldPop project at the University of Southampton (WorldPop (2018)), researchers have used information from satellite imagery to identify areas of settlements, and combined this information with census data to obtain highly granular population

density estimates across Africa (Linard et al. (2012)). These results have then been combined with data on age- and sex-distributions from censuses (or more recent surveys) to map the distribution of populations by age and sex. While the fine-grained resolution of this work is impressive, there are two main drawbacks in using these estimates as denominators to track health indicators over time. Firstly, the population age distribution is based on observations from the most recent census, or survey data if a census is not available. Little attention is paid to how age distributions across regions change over time. In addition, it is unclear how uncertain estimates in a particular region may be, and how that uncertainty varies over geographic space and time. Understanding uncertainty around population sizes and flows is essential in quantifying overall uncertainty about a country's progress through time.

The methodology proposed in this paper incorporates a cohort component projection model into a Bayesian hierarchical framework to understand changes in population structures over time. It allows estimates to be driven by available data and for uncertainty to be incorporated around estimates and projections. The approach has similarities with methodologies described in Wheldon et al. (2013) (but with a focus on subnational estimation) and in Bryant and Graham (2013) (but with a focus on data-sparse situations).

3 Data

We aim to estimate female population counts for ages 15-49 for subnational areas that are the second administrative level down. This data description focuses on Kenya, for which the model is applied in later sections. However, the data and methods are more broadly applicable to other countries that have similar data available. Inputs used to obtain estimates come from two main sources: censuses, and national population and mortality estimates from the 2017 World Population Prospects. These data sources are outlined in the following sections. In addition, we discuss other data sources that were considered as potential inputs, but were not used at this stage of the project.

3.1 Census data

Data inputs on subnational population counts and internal migration flows come from national censuses. The census data are available through Integrated Public Use Microdata Series (IPUMS) International (Minnesota Population Center 2017). IPUMS-International contains samples of microdata for 305 censuses over 85 different countries. The data are harmonized to create temporally stable variables and geographic boundaries. The harmonization over time is useful in modeling population over time without having to deal with apportioning populations across changing geographic boundaries. The majority of countries of interest have relatively recent censuses available through IPUMS-International. For example, Kenya has decennial censuses available from 1969 to 2009.

3.1.1 Subnational population counts

For inputs to the model, female population counts by single year of age for ages 15-49 and subnational administrative region are obtained directly from the IPUMS-International microdata. As these data are samples (most commonly 10%), the microdata are multiplied by the person weights to obtain counts by age and area.[1] For Kenya, the first administrative units are provinces, and the second administrative units in IPUMS are districts. There are eight provinces and 35 districts. The districts represent slightly larger groups than the 47 Kenyan counties, which are harmonized and temporally stable (IPUMS (2018)). Provinces and districts are illustrated below in Fig 1.

Note that the raw data may suffer from age heaping, where population counts display a preference for ages ending in 0 or 5. Following standard practice (e.g. UNPD (2017b)) the raw data are smoothed to adjust for

[1] The sampling error introduced by considering sampled microdata is accounted for in the data model, refer to the Methods section for details.

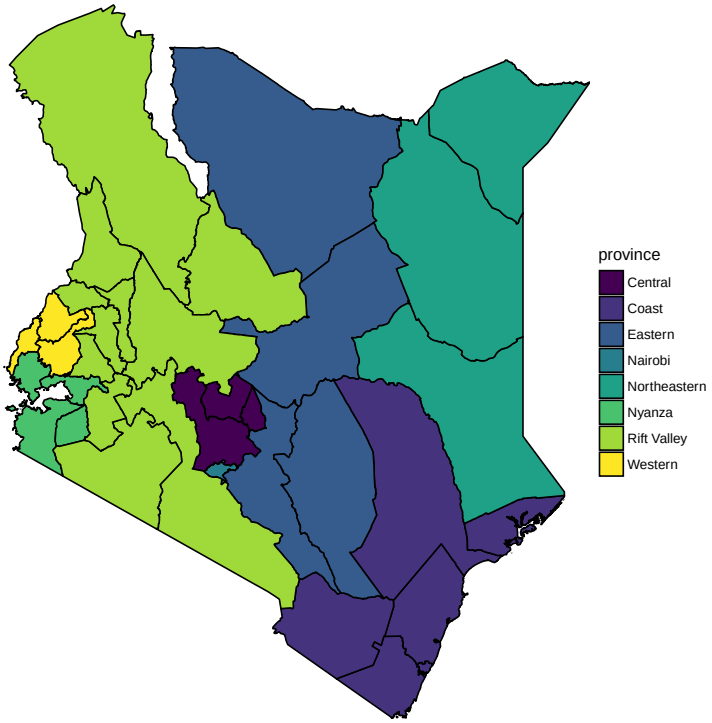


Figure 1: Map of Kenya provinces, showing IPUMS harmonized districts.

problems with age heaping.[2]

3.1.2 Net-migration counts

Information on internal net-migration is also obtained from national censuses. Net-migration is the difference between the number of in-migrants and out-migrants in an area. Thus, it can either be positive, if more people move to an area than leave, or negative, if the opposite is true. For the case of Kenya, net-migration counts can be derived using information from two questions:

- A person's location one year ago (IPUMS harmonized variable `MIGRATE1`). The information is provided in the form of categories, i.e. same major administrative unit, different minor administrative unit; different major and minor administrative unit; abroad. Migrants are limited to those who were in a different minor or major administrative unit one year ago.
- A person's district of residence one year ago (for Kenya, this is the variable `MIGKE`).

Given these two variables, we can calculate the net-migration by age for a particular district in the year before the census. For example, to get counts for net-migration for Nairobi in 2008, use information from the 2009 census and:

- calculate the number of in-migrants by summing residents in Nairobi who indicated they were in a different administrative unit one year ago;
- calculate the number of out-migrants by summing residents in other districts who indicated they lived in Nairobi one year ago; and
- take the difference between the in-migrants and out-migrants.

We calculated net-migration in this way by age, for each district and each census.

[2]We used a smoothing spline ('smooth.spline' in R) to smooth population counts across age. In future work, we aim to formulate a Bayesian method to adjust for age heaping, which incorporates a penalized splines regression with known digit preferences in age reporting.

3.2 National estimates from WPP

The World Population Prospects (WPP) are the official population estimates and projections produced by the United Nations. WPP is revised every two years, with the latest revision being in 2017 (UNPD 2017a). WPP estimates are produced using a combination of census and survey data, and demographic and statistical methods. Both population counts and mortality estimates from WPP are used in the model.

It is important to ensure that the sum of population estimates at the regional level agrees with published estimates at the national level. National population counts produced by WPP are used as a constraint in the model. The WPP models populations of five-year age groups every five years from 1950-2100.

National mortality estimates produced by WPP are used as the basis of a mortality model for patterns at the regional level. WPP uses the relationship between infant mortality and the probability of dying between ages 15 and 60, i.e. ${}_{45}q_{15}$, to estimate a life table based on Coale-Demeny Model Life Tables (UNPD 2017b). We use estimates of the probability of dying between ages x and $x+5$, ${}_5q_x$. As we estimate population by single year age, we convert these five year probabilities to probabilities of death between x and $x+1$, ${}_1q_x$, using linear interpolation.

3.3 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.

3.3.1 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 through the 'maternal mortality module'. 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 (World Health Organization (WHO) 2016a). However, the DHS does not ask the location of residents of the siblings who died, thus the data is difficult to use at the subnational level. It could be assumed that there is no migration between areas and use sibling mortality as a basis of estimates of mortality in the area of the interview respondent. However, this would confound estimates of mortality and migration parameters in the model, and given death counts by district can be quite small and uncertain, the sibship mortality information is not currently included. It should be noted that mortality information from the DHS could potentially be used at the national level, in replacement of (or as well as) the WPP estimates. This would avoid the reliance on WPP mortality estimates, which themselves are modeled using Coale-Demeney model life tables (UNPD (2017b)). Future work will investigate the use of DHS mortality data as a potential input at the national level.

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.

3.3.2 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 two most recent censuses (1999 and 2009). 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.[3] 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.

4 Model

Our aim is to obtain estimates of the number of women aged 15-49 by region and year for a certain time period. Let $\eta_{r,t}$ be the (true) population of women of reproductive age in region r at time t. In our modeling framework we consider age-groups of women from a cohort perspective and project population counts in each age through time. In particular, we will estimate the number of women at each age a and cohort c in region r, $\eta_{r,a,c}$, and then sum the relevant ages (15-49) and cohorts to obtain $\eta_{r,t}$:

$$\eta_{r,t} = \sum_{a;c[t]} \eta_{r,a,c} \tag{2}$$

We chose to model from a cohort perspective because this allows some demographic structure about mortality and migration trends across age and cohorts to be built into the model. In particular, we build up from a cohort component framework (Leslie (1945), Wachter (2014)), which relates population counts to mortality and migration patterns. Mortality and migration are then modeled in a hierarchical framework. Note that as the youngest age that we are interested in is age 15, we do not consider fertility or births as part of the model.

This section describes the model in detail. For a summary of the model set-up, see Appendix A.

4.1 Model for true population

We model the true number of women at age a in cohort c and region r as

$$\eta_{r,a,c} = \eta^*_{r,a,c} \cdot \varepsilon_\eta \tag{3}$$

where η^*_{rac} is the expected number of women and ε_η is some distortions around the expected level. On the log scale this is equivalent to

$$\log \eta_{r,a,c} = \log \eta^*_{r,a,c} + \log \varepsilon_\eta \tag{4}$$

We assume $E(\log \varepsilon_\eta) = 0$ and $Var(\log \varepsilon_\eta) = \sigma^2_\eta$. The expected number of women at age a in cohort c and region r is based on a cohort component method,

$$\eta^*_{r,a,c} = \eta^*_{r,a-1,c} \cdot \rho_{r,a-1,c} + \phi_{r,a-1,c} \tag{5}$$

[3]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.

that is, the number in age group a is the number in the previous age $a-1$ in that cohort times the survivorship multiplier $\rho_{r,a-1,c}$, plus net migration $\phi_{r,a-1,c}$.

4.2 Data model for observed population counts

Let $y_{r,a,c}$ be the observation of population in region r, age a and cohort c.[4] We assume that:

$$y_{r,a,c} \sim N(\eta_{r,ac}, s^2_{rac}) \tag{6}$$

That is, the population count we observed is distributed around the true population with some error. The s^2_{rac} captures sampling error in the data observations. As we are using the IPUMS 10% sample microdata, the sampling variance is derived from assuming the observed population is a draw from a binomial distribution with $p = 0.1$ and n equal to the observed population y_{rac}.

4.3 Priors on first year and age

As a consequence of the cohort component setup, the population at a particular age is estimated using the population at the previous age in the same cohort. As such we need to set priors on the populations at the initial age of estimation (age 15) and first year (1979). Define $b_{r,c}$ as the mean of the prior on the first age in region r and cohort c and $d_{r,f-a}$ as the mean of the prior on the population in the first period f, which corresponds to cohort $f - a$. To obtain these values:

- The proportion of the total population in each age and region is calculated for each census year.
- These proportions are linearly interpolated across the whole estimation period, and applied to the WPP estimate of the national population in each year.
- b_{rc} is then the estimated population of 15 year olds in each region and cohort; and $d_{r,f-a}$ is are the estimated populations by age from the first year.

We then assume priors of the form:

$$\log \eta_{r,1,c} \sim N(\log b_{r,c}, 1) \tag{7}$$
$$\log \eta_{r,a,(f-a)} \sim N(\log d_{r,a}, 1) \tag{8}$$

Choosing a variance equal to 1 means these priors are relatively uninformative.

4.4 Mortality

To estimate the survivorship multipliers, $\rho_{r,a-1,c}$, we want to estimate the expected conditional probability of survival given age a and cohort c. This is equal to the complement of the probability of dying in the age interval, i.e.

$$\rho_{r,a-1,c} = 1 - q_{r,a-1,c} \tag{9}$$

where $q_{r,a-1,c}$ is the probability of dying between ages $a-1$ and a.

We use information about mortality trends at the national level as the basis for a mortality model at the subnational level. A semi-parametric model is used to capture shape of national mortality through age and time, while allowing for differences by region. In particular, we model regional mortality on the logit scale as

[4] Note that as discussed in the Data section, the observed counts have been adjusted to account for age-heaping.

$$\text{logit } q_{r,a,c} = \text{logit } \bar{q}_a + \beta_{1,r,c} \cdot Y_{1,a} + \beta_{2,r,c} \cdot Y_{2,a} \tag{10}$$

where logit $\bar{q}_a$ 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.

Principal components create an underlying structure of the model in which regularities in age patterns of human mortality can be expressed. Many different kinds of shapes of mortality curves can be expressed as a combination of the components. Incorporating more than one principal component allows for greater flexibility in the underlying shape of the mortality age schedule. Using SVD for demographic modeling and forecasting first gained popularity after Lee and Carter used the technique as a basis for forecasting US mortality rates (Lee and Carter (1992)). More recently, SVD has become increasingly used in demographic modeling, in both fertility and mortality settings (Schmertmann et al. (2014); Clark (2016); Alexander, Zagheni, and Barbieri (2017)).

Principal components were obtained from a decomposition on a matrix which contains a set of standard mortality curves. As discussed in the Data section, we used national Kenyan life tables published in the World Population Prospects 2017 (UNPD 2017a). The mean mortality schedule and the first two principal components for Kenyan national mortality curves between ages 15-49 from 1950–2020 are shown in Fig. 2. We used constrained principal components computation to ensure all components were non-negative.[5] This was done to ensure HIV/AIDS mortality would affect each age in the same direction.

The mean logit mortality schedule shows a standard age-specific mortality curve, with mortality increasing over age. The first two principal components have demographic interpretations. The first shows the average contribution of each age to mortality improvement over time. This interpretation is similar to the b_x term in a Lee-Carter model (Lee and Carter 1992). For the case of Kenya, the second principal component most likely represents the relative effect of HIV/AIDS mortality by age.

The principal component coefficients for each region and year, $\beta_{d,r,c}$ (for $d = 1, 2$) are assumed to be a draw from a national distribution:

$$\beta_{d,r,c} \sim N(\mu_{d,c}, \sigma^2_{d,c}) \tag{11}$$

This hierarchical structure allows information about mortality trends to be pooled across regions. The mean coefficient parameters $\mu_{d,c}$ are modeled as a random walk process:

$$\Delta\mu_{d,c} \sim N(0, \sigma^2_d) \tag{12}$$

4.5 Migration

In addition to mortality, the model for the true population (Eq. 5) also contains a migration term, ϕ_{rac}. As this is capturing net-migration, it can either be positive or negative. Allowing a migration term for each age in each year and region requires many parameters to be estimated, with very little information available. However, when looking at the available data on migration for Kenya, there were clear patterns by age and region, allowing for a simpler migration model with much fewer parameters to be proposed.

Figure 3 shows net migration as a proportion of total population for each of the districts in Kenya based on the available census data. This chart suggests that, for the majority of districts, net migration is negligible relative to overall population size.

[5] Non-negative principal components were calculated in R using the `nsprcomp` package.

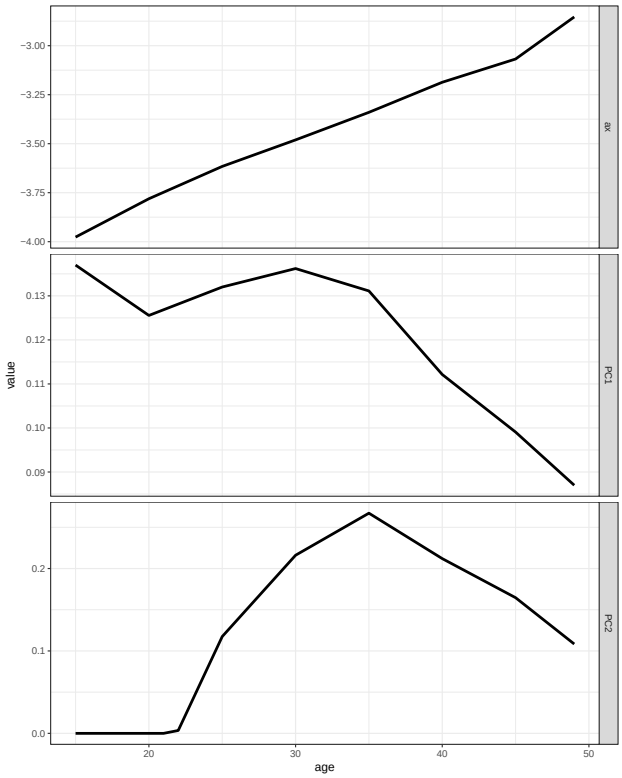


Figure 2: Mean logit mortality and first two principal components

As a consequence, for modeling purposes we assume net migration is zero for 23 of the 35 districts. For the remaining 12 districts, listed in the table below, net migration is modeled.

District	Areas covered
404001001	Nairobi, Westlands
404004003	Embu, Kangundo, Kibwezi, Machakos, Makueni, Mbeere, Mbooni, Mwala, Nzaui, Yatta
404004004	Kitui North, Kitui South (Mutomo), Kyuso, Mwingi
404006001	Bondo, Rarieda, Siaya
404002002	Nyeri
404006004	Borabu, Gucha, Kisii, Manga, Masaba, Nyamira
404007006	Eldoret East, Eldoret West, Wareng
404007009	Kaijiado, Loitoktok, Molo, Naivasha, Nakuru
404008001	Butere, Emuhaya, Hamisi, Kakamega, Lugari, Mumias, Vihiga
404008002	Bungoma, Mt. Elgon
404008003	Bunyala, Busia, Samia, Teso
404003001	Kilindini

Figure 4 shows the age distribution of net migration from census data in the districts where migration is modeled. These data suggest that the age distributions over time are fairly stable. As such, for model simplicity, we assume a stable age distribution of migration across time.

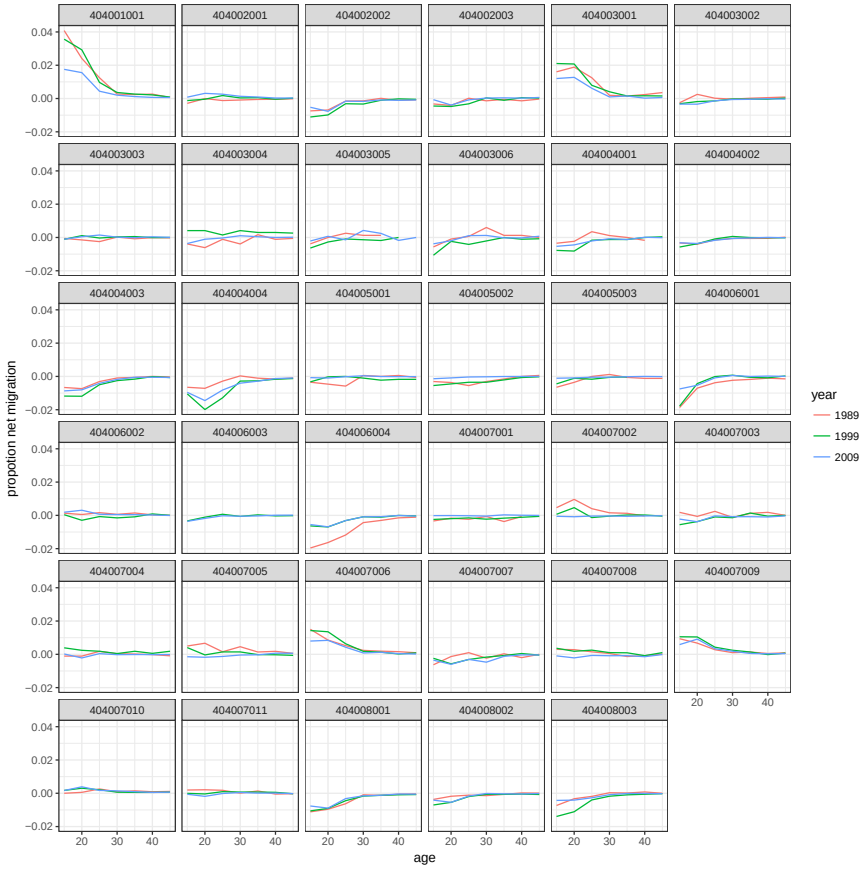


Figure 3: Net migration as a proportion of population, by district and year

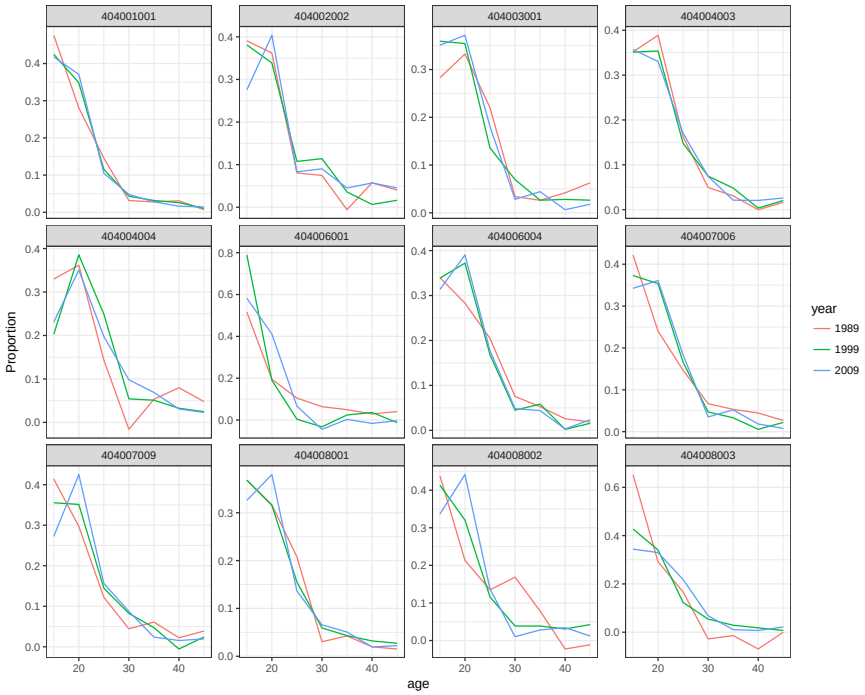


Figure 4: Net migration by age as a proportion of total net migration

4.5.1 Model for net migration

The net migration in region r, age a and cohort c, $\phi_{r,a,c}$ is modeled as

$$\phi_{r,a,c} = \begin{cases} \eta_{r,c} \cdot \pi_{r,c} \cdot A_{r,a}, & \text{if region is a migration region} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

where

- $\eta_{r,c}$ is the total population (summed across age) in region r and cohort c
- $\pi_{r,c}$ is the proportion of total population that is net-migration
- $A_{r,a}$ is the proportion of total net-migration at age a for region r. This is derived from the data; taking the average of smoothed age distributions across all census years (i.e. the average of data shown in Fig. 4).

The 'true' proportion of the total population that is net-migration, π_{rc} is estimated, partially informed by census data inputs. In particular, it is assumed that the ith observed proportions, $P_{rc[i]}$, are distributed around the true proportions, with some error:

$$P_{r,c[i]} \sim N(\pi_{r,c}, \sigma_P^2) \tag{14}$$

The migration proportion for each region and cohort, π_{rc} is modeled as a random walk, constrained to be within bounds of $[-0.2, 0.2]$, i.e.

$$\pi_{r,c} \sim N(\pi_{r,c-1}, \sigma_\pi^2) T[-0.2, 0.2] \tag{15}$$

These truncation bounds were chosen based on looking at reasonable upper and lower bounds from the data.

4.6 National Constraint

An important part of estimating subnational populations is to ensure the sum across all regions is consistent with previously published national population estimates. In particular, we would like to ensure that population counts for each five-year age group are consistent with the national population estimates published as part of the WPP (UNPD 2017a). The WPP models populations of five-year age groups at the mid-point of every five years. As such, we constrain the population in each five year age group to be within bounds that are approximately 90% and 110% of the relevant WPP estimate, and this constraint is implemented every five years for WPP years e.g. 1982, 1987, ..., 2017. These lower and upper bounds are estimated within the model:

$$L_{g,y} < \sum_{a[g],r} \eta_{a,y} \leq U_{g,y} \tag{16}$$

$$\log L_{g,y} \sim N(\log 0.9 WPP_{g,y}, 0.1) T(, \log WPP_{g,y}) \tag{17}$$

$$\log U_{g,y} \sim N(\log 1.1 WPP_{g,y}, 0.1) T(\log WPP_{g,y},) \tag{18}$$

where

- $L_{g,y}$ and $U_{g,y}$ are the lower and upper bounds on the national population in age group g and WPP year y.
- $WPP_{g,y}$ refers to the WPP estimate of the national population in age group g and WPP year y.

Note that $T(,)$ refers to truncation of the distribution with particular bounds. The lower bound $L_{g,y}$ is not left-truncated but is right-truncated to be no more than $\log WPP_{g,y}$. For the upper bound, the opposite is true.

4.7 Projection

While the modeling framework focuses on reconstruction of past WRA, projection of populations can be incorporated. The model set-up allows trends in population counts by age and region to be projected into the future. The mean coefficients $\mu_{d,c}$ can be projected forward according to the setup defined in Eq. 12 and these can then be used in combination with the principal components to obtain projections and uncertainty for population counts. The WPP produces projections of national populations up to 2050, and these can be incorporated into the model to include a constraint on the sum of the subnational projections.

In particular, to project the number of women of age a in region r in cohort $c+1$:

1. draw values for $\mu_{d,c+1}$ based on Eq. 12 and the estimate for σ_d;
2. use these values to calculate the probability of death (based on Eq. 10) and the corresponding mortality multiplier $\rho_{r,a,c+1}$;
3. calculate values for $\phi_{r,a,c+1}$ based on Eq. 13; and
4. use these values to calculate value for $\eta_{r,a,c+1}$ based on Eq. 5.

Note that there are no restrictions on the projections of the principal component coefficients. However, given that the second principal component appears to relate to HIV/AIDS mortality, we would expect coefficients related to this dimension to eventually reach zero. Future work will focus on building plausible projections for these coefficients.

4.8 Computation

The model was fitted in a Bayesian framework using the statistical software R. Samples were taken from the posterior distributions of the parameters via a Markov Chain Monte Carlo (MCMC) algorithm. This was performed using JAGS software (Plummer 2003). Standard diagnostic checks using trace plots and the Gelman and Rubin diagnostic (Gelman and Rubin 1992) were used to check convergence.

Best estimates of all parameters of interest were taken to be the median of the relevant posterior samples. The 95% Bayesian credible intervals were calculated by finding the 2.5% and 97.5% quantiles of the posterior samples.

5 Results

We fitted the model to 35 Kenyan districts over the period 1979-2020. As discussed in the Model section, we obtained estimates of population at every age between 15-49, as well as parameters associated with mortality and migration. This section highlights some key features of the results.

5.1 Population over age, period and cohorts

Fig. 5 shows the WRA population by province in 1979-2020. The black line and associated shaded area are the model estimates and associated 95% credible intervals. The red dots and error bars are the data from decennial censuses and associated sampling error. Populations of WRA are increasing in every province, with the two largest provinces being Nairobi and Rift Valley. While Northeastern is the smallest province by population size, the growth rate is relatively rapid. This is likely due to the relatively high fertility rates in this province (Westoff and Cross (2006); Kenya National Bureau of Statistics (2015)), whereas rapid population increases in Nairobi are driven by in-migration. Fig. 6 shows population increases are concentrated in districts surrounding Nairobi.

Fig. 7 decomposes the population of WRA into the proportion by age for each year for each province. For most provinces, the proportion of the total WRA decreases by age. The notable exception is Nairobi, where

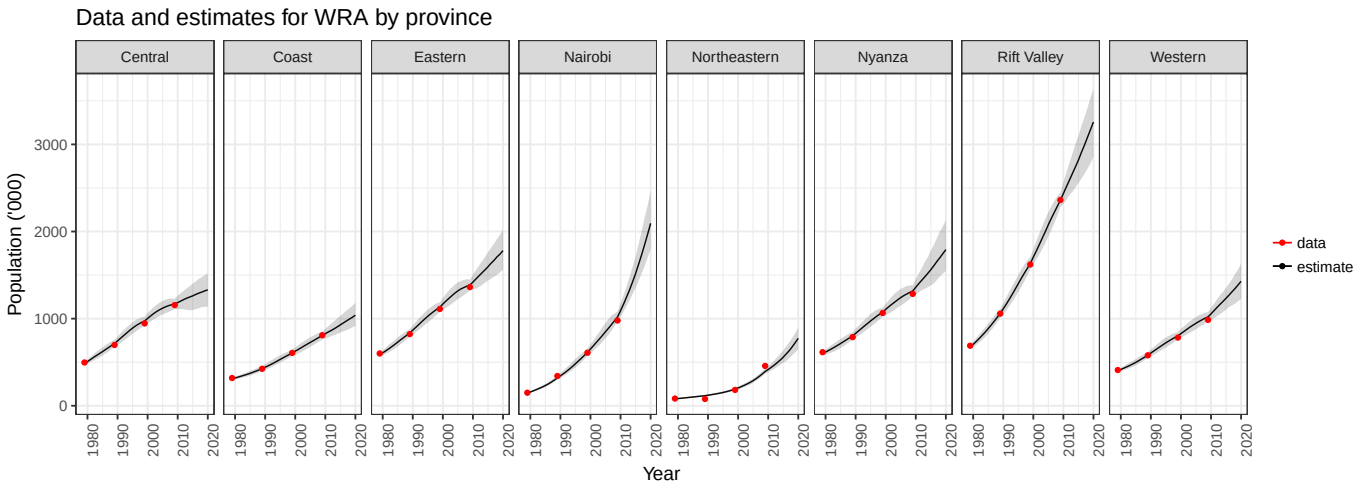


Figure 5: Data and estimates of WRA by province, Kenya, 1979-2020.

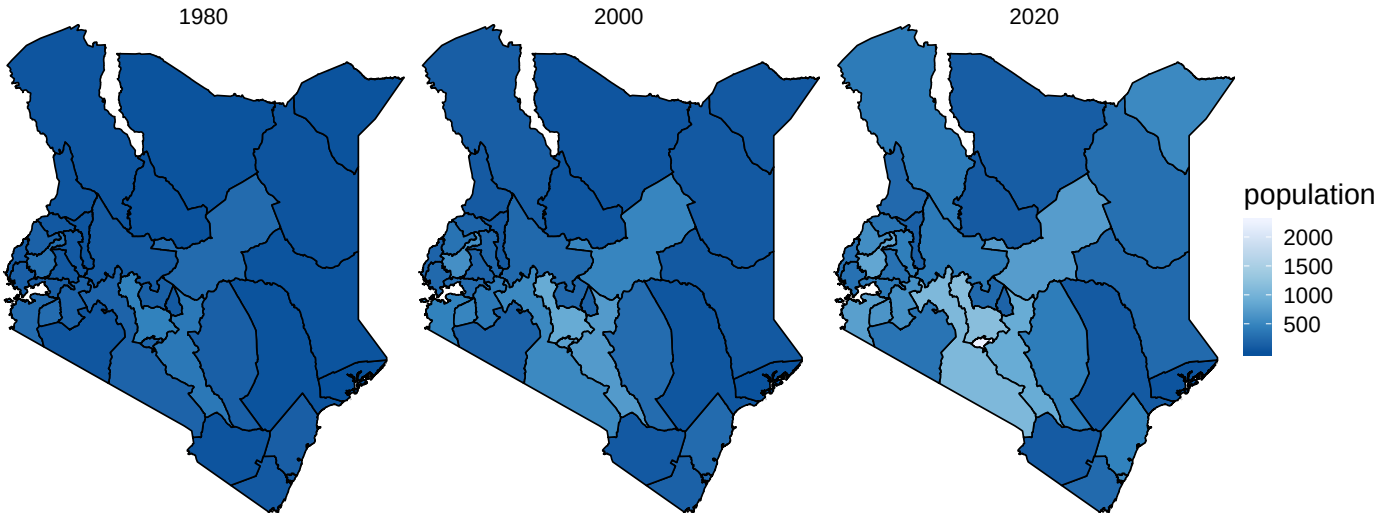


Figure 6: Estimates of population ('000) by district: 1980, 2000 and 2020.

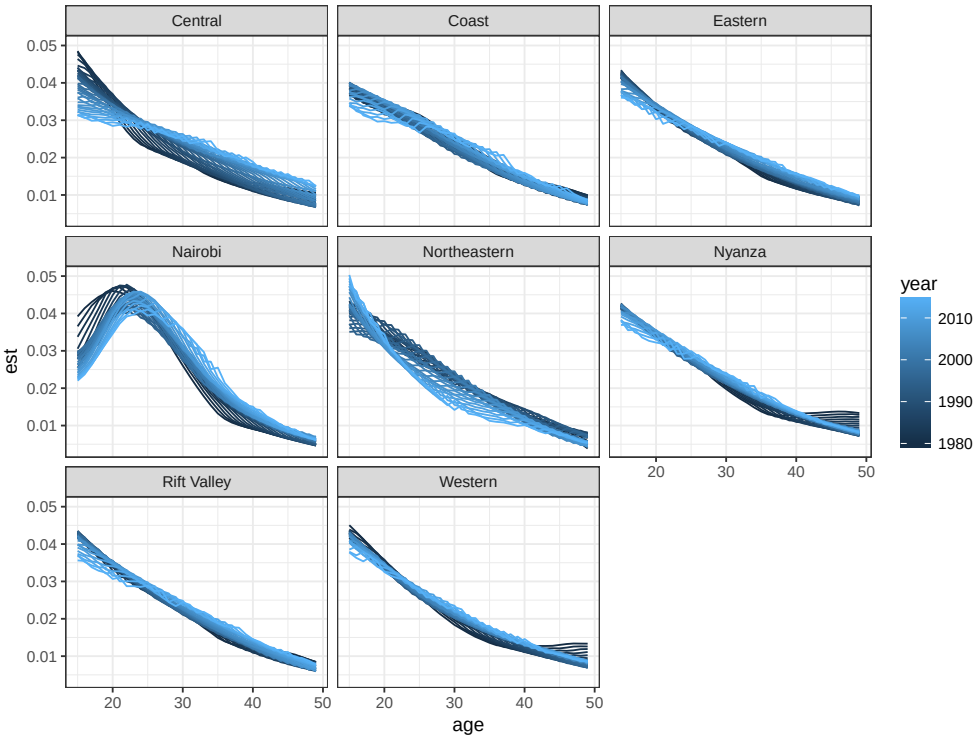


Figure 7: Estimates of age structure by province, Kenya, 1979-2020.

net in-migration at younger working ages causes the age distribution to peak between 20-30. The Nairobi age distribution is becoming increasingly older, possibly due to the declining fertility rates. Indeed, for most other provinces the slope of the age distribution is decreasing over time, leading to an older population of WRA. The exception is Northeastern, where fertility rates are still very high (Kenya National Bureau of Statistics (2015)).

The populations of WRA at each age can also be visualized in terms of cohorts. Fig. 8 shows the populations by age for each cohort in four different districts: Embu, where there is net out-migration, Kaijiado, which has modest in-migration, Kiambu, where there is zero net-migration, and Nairobi, where there is high in-migration. The populations at each age are broadly increasing across cohort, although the effect of high HIV/AIDS mortality can be seen as the 'kinks' in population affecting each cohort at different ages.

5.2 Comparison with WPP national estimates

Within the model framework, the sum of the subnational population estimates is constrained to be within 10% of the WPP estimates. Fig. 9 shows the estimated national total by year, calculated by summing across areas (shown in black) compared to the WPP estimates for Kenya over the same period. The median estimates for the total population are very similar to WPP. Uncertainty around the national total increases markedly in the projection period, beginning 2009.

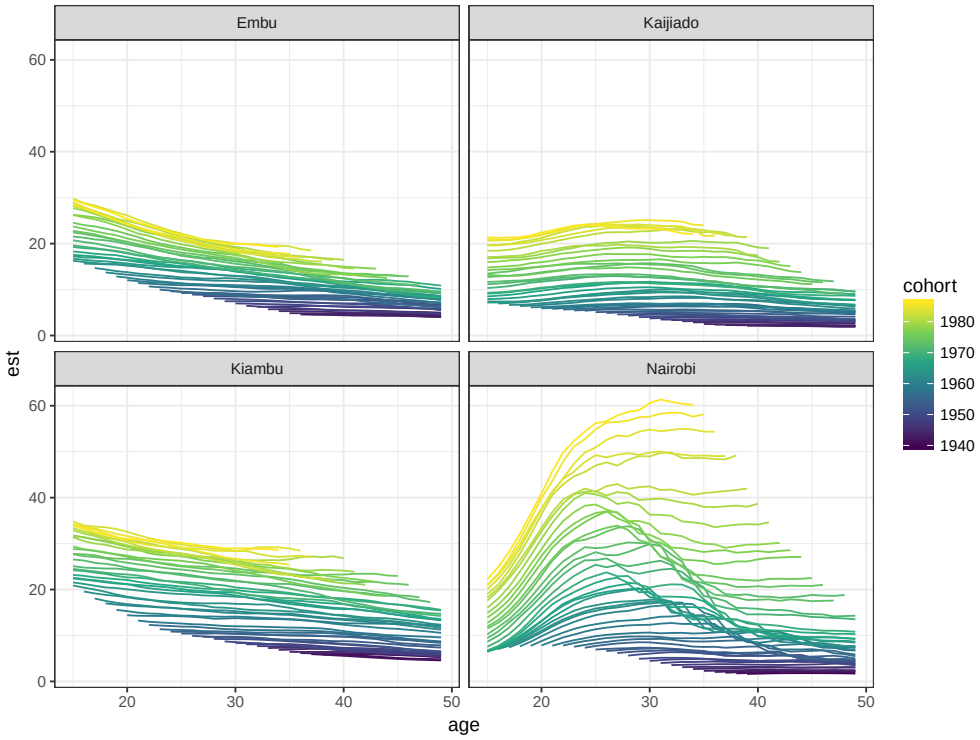


Figure 8: Population estimates by age and cohort, four districts in Kenya.

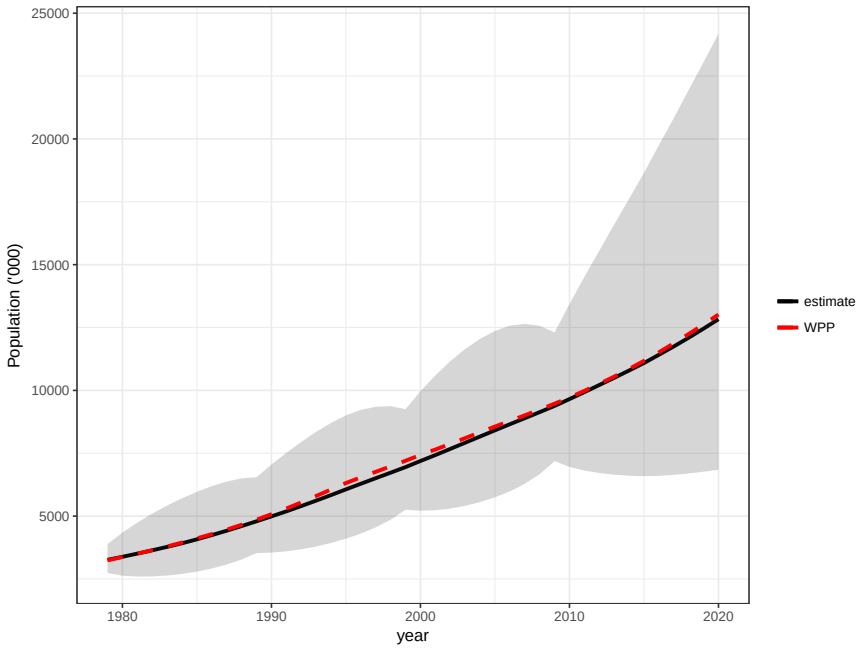
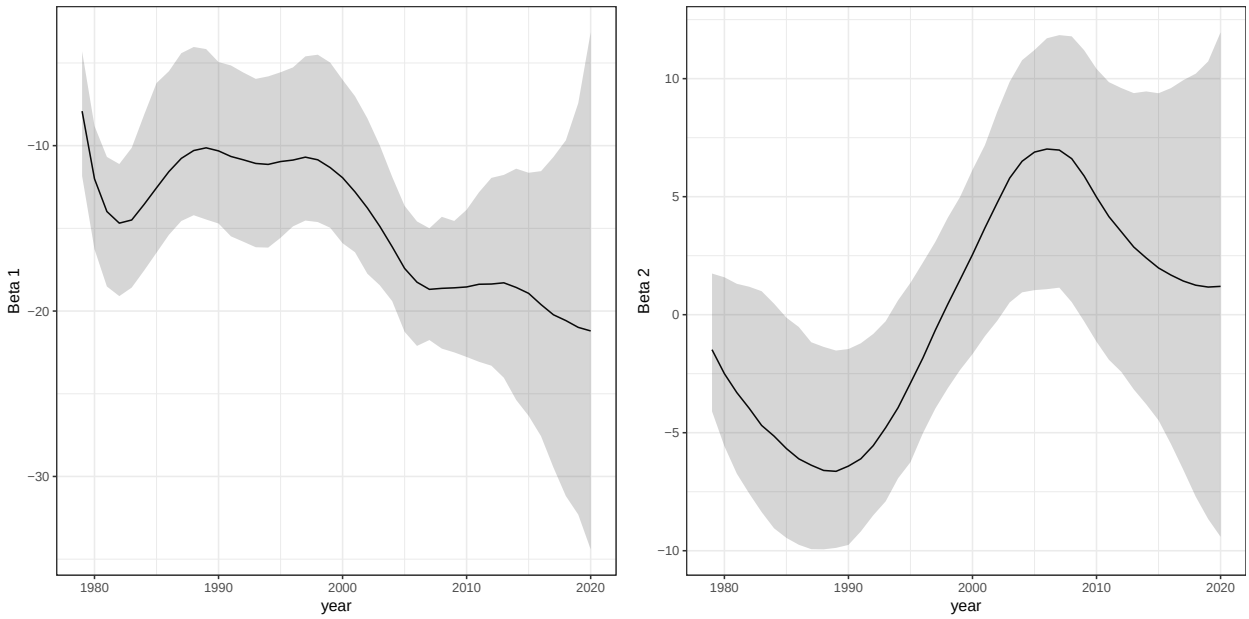


Figure 9: Comparison of WPP and national estimates.

5.3 Mortality change

Mortality patterns are estimated with the model as a linear combination (on the logit scale) of a mean mortality schedule plus two principal components, which respectively broadly capture age-specific contributions to overall mortality change, and HIV/AIDS mortality. The figures below show estimates and 95% credible intervals for the national mean coefficients, i.e. $\mu_{d,c}$ over time (refer to Eq. 11). Results for the first principal component, $\mu_{1,c}$ suggest that overall mortality is generally declining over time, with a plateau in progress during the 1990s. The coefficient on the second principal component, $\mu_{2,c}$, shows an increase in HIV/AIDS mortality during the 1990s and 2000s, peaking around 2005, before declining through the projection period.



Region-specific coefficients of mortality showed very similar patterns of change over the period. The variance of the β_1's around the μ_1 was estimated to be 9.5 [95% CI: 5,8, 12.7] while the variance of the β_2's around the μ_2 was estimated to be 1.8 [95% CI: 0.3, 5.3]. Broadly, there was some evidence to suggest overall mortality levels were highest in Nyanza and Western, and lowest in Rift Valley and Coast (although differences were not statistically significant). During the years where HIV/AIDS mortality peaked, it was highest in Nairobi and lowest in Northeastern.

6 Discussion

In this paper we proposed a Bayesian cohort component projection framework to estimate the population of women of reproductive age by subnational regions. The model uses information on population and migration counts from censuses, as well as mortality patterns from national schedules, to reconstruct WRA populations based on cohorts moving through time. The modeling framework also naturally extends to allow projection of populations. The model ensures the national WRA populations implied by the sum of subnational areas agree with pre-published UN estimates.

The model was used to estimate and project WRA populations for 35 districts in Kenya over the period 1979-2020. Results suggested continued growth of WRA populations in all districts, and accelerated growth in particular in areas such as Nairobi and Northeastern. In general, the average age of WRA has increased over time. The mortality component of the modeling framework highlighted the stagnating progress through the 1990s and 2000s, largely due to HIV/AIDS, but more recent mortality declines. The model requires inputs from national censuses and WPP estimates, which are available for the majority of countries. Thus, while the

model was tested on estimation in Kenya, the methodology is applicable to a wide range of countries with very little alterations.

There are several advantages and contributions of this modeling framework to the estimation of subnational populations. The model is governed by processes of mortality and migration, tracking cohorts of women as they move through time. This has advantages over more aggregate techniques such as interpolation and extrapolation, because it ensures that populations of women at each age will always make sense from a demographic point of view (for example, a population of women at age a can never be more than the same cohort at age $a-1$, above what is implied by net migration). In addition, this process takes into account intercensal events such as trends in HIV/AIDS mortality.

Secondly, the modeling framework proposes a parsimonious model for internal net-migration across subnational areas. In cohort component models, it is often the case that migration components are assumed to be negligible or considered to just be the residual once mortality has been taken into account. Very little data usually exists on migration patterns, and estimation of all migration components by age, region and year becomes very intensive. After observing key patterns in the data, we proposed a net-migration model which separates migration patterns into independent age and time components. The result is an age-specific net migration model with parameters that are more easily identifiable. However, while the current migration model works well for the case of Kenya, it may be too restrictive for other countries. The migration component could be modeled more flexibly by including changes in age structure over time, for example. Modeling decisions should be informed by the data that are available.

The incorporation of a cohort component projection model into a probabilistic setting allows for different sources of uncertainty, such as sampling and non-sampling error, to be included into the modeling process. The Bayesian hierarchical framework allows information from different data sources to be consolidated without the need for post-estimation redistribution changes as is often the case with subnational population estimation (Swanson and Tayman (2012)). In addition, it allows for increased flexibility in modeling population processes compared to traditional deterministic techniques, while still keeping the basis of an underlying demographic process.

6.1 Future work

The proposed model produced promising initial results when applied to the reconstruction and projection of WRA populations in Kenya. Future work will focus on investigating the possibility of including additional data sources to inform subnational mortality patterns. Results of the mortality parameter estimation suggests the mortality component in the model would benefit from additional information and constraints. For example, the credible intervals around the mean coefficients on the principal components are relatively large, particularly for the projection period. This suggests that there are many different combinations of the mean mortality schedule and two principal components that lead to reasonable population change over time. Given what we know about overall mortality change in Kenya, and in particular, decreases in HIV/AIDS mortality (Kenya National Bureau of Statistics (2015)), constraints could be put on mortality parameters in projections; for example, the HIV/AIDS mortality component could be assumed to decline to an eventual level of zero. In addition, there is very little variation in mortality estimated across regions, even though mortality variation across urban and rural areas in Kenya has been documented elsewhere (Kenya National Bureau of Statistics (2015)). As discussed in the Data section, we chose to omit subnational mortality data available from sibling histories, because of issues surrounding the location of siblings at death, in combination with well-established issues of sibling methods, especially when sample sizes are small (Masquelier (2013)). However, future work will focus on assessing the usability of this data at both the national and subnational level, and of different methods to utilize this information, including both sibling and network survival methods (Feehan, Mahy, and Salganik (2017)).

Additionally, we plan to investigate other forms of age heaping adjustments. As discussed in the Data section, the input to the model is smooth census counts. However, it would be possible to incorporate this adjustment into the same modeling framework, thereby incorporating the uncertainty associated with smoothing into the final estimates. Building on work by Camarda, Eilers, and Gampe (2008), one approach would be to model

the underlying true distribution of age counts using penalized splines regression. This is then adjusted based on a series of transition probabilities that represent the probability of age-misreporting.

Finally, the modeling framework does not consider fertility rates in the population accounting equation. Although we are not explicitly estimating female births, WRA populations are implicitly affected by changes in fertility rates over time, as the size of the birth cohort partially determines the size of the cohort at age 15 and above. Future work will investigate the relative trade-off between the additional information gained by incorporating fertility trends, compared to the drawbacks of including potentially biased, poor-quality births data and increasing model and estimation complexity.

A Complete model

The current model is

$$
\begin{aligned}
y_{r,a,c} &\sim N(\eta_{r,a,c}, s^2_{r,a,c}) \\
\eta^*_{r,a,c} &= \eta^*_{r,a-1,c} \cdot \rho_{r,a-1,c} + \phi_{r,a-1,c} \\
\log \eta_{r,a,c} &\sim N(\log \eta^*_{rac}, \sigma^2_\eta) \\
\log \eta_{r,1,c} &\sim N(\log b_{r,c}, 1) \\
\log \eta_{r,a,(f-a)} &\sim N(\log d_{r,a}, 1) \\
\rho_{r,a,c} &= (1 - q_{a,p[c]}) \\
\text{logit } q_{r,a,c} = \text{logit } \bar{q}_a + \beta_{1,r,c} \cdot Y_{1,a} + \beta_{2,r,c} \cdot Y_{2,a} & \\
\beta_{d,r,c} \sim N(\mu_{d,c}, \sigma^2_{d,c}) & \\
\Delta\mu_{d,c} \sim N(0, \sigma^2_d) & \\
\phi_{r,a,c} &= \begin{cases} \eta_{r,c} \cdot \pi_{r,c} \cdot A_{r,a}, & \text{if region is a migration region} \\ 0, & \text{otherwise.} \end{cases} \\
\eta_{r,c} &= \sum_a \eta_{r,a,c} \\
P_{r,c} &\sim N(\pi_{r,c}, \sigma^2_P) \\
\pi_{r,c} \sim N(\pi_{r,c-1}, \sigma^2_\pi) T[-0.2, 0.2] L_{g,y} < & \textstyle\sum_{a,r} \eta_{g,y} \leq U, \\
\log L_{g,y} &\sim N(\log 0.9 WPP_{g,y}, 0.1) T(, \log WPP_{g,y}) \\
\log U_{g,y} &\sim N(\log 1.1 WPP_{g,y}, 0.1) T(\log WPP_{g,y},)
\end{aligned}
$$

Explanation of symbols:

- $y_{r,a,c}$ is the observation of population in region r, age a and cohort c. This has been smoothed using the built in splines smoother in R.
- $\eta_{r,a,c}$ is the true population in region r, age a and cohort c
- $b_{r,c}$ is the prior on the first age group in region r and cohort c. This is derived from multiplying the $WPP_{1,c}$ by the proportion of people in age group 1 in region r in cohort c. The proportion is derived from linearly interpolating the census proportions.
- $d_{r,c}$ is the prior on the population in the first period $f - c$. This is derived from multiplying the $WPP_{a,(f-a)}$ by the proportion of people in each age group in region r in cohort c.
- $\rho_{r,a,c}$ is a mortality multiplier: the expected conditional probability of survival given age a and cohort c. This is equal to the complement of the probability of dying in the age interval.
- ${q}_a is the mean age-specific mortality schedule of the standard logit mortality curves
- The β's are the coefficients associated with the principal components
- The Y's are the first and second principal component of the demeaned standard logit mortality curves
- $\phi_{r,a,c}$ is the net-migration component. This is set to zero if region is deemed to have negligible migration.
- $\pi_{r,c}$ is the proportion of total population that is net-migration
- $A_{r,a}$ is the proportion of total net-migration at age a for region r. This is derived from the data; taking the average of smoothed age distributions in census years.
- $P_{r,c}$ is the observation of proportion of total population that is net-migration in region r and cohort c.
- $L_{g,y}$ and $U_{g,y}$ are the lower and upper bounds on the national population in age group g and WPP year y.
- $WPP_{g,y}$ refers to the World Population Prospects estimate of the national population in age group g and WPP year y.

References

Alexander, Monica, and Leontine Alkema. 2018. "Global Estimation of Neonatal Mortality Using a Bayesian Hierarchical Splines Regression Model." *Demographic Research* 38: 335–72.

Alexander, Monica, Emilio Zagheni, and Magali Barbieri. 2017. "A Flexible Bayesian Model for Estimating Subnational Mortality." *Demography* 54 (6). Springer: 2025–41.

Alkema, Leontine, and Jin Rou New. 2014. "Global Estimation of Child Mortality Using a Bayesian B-Spline Bias-Reduction Model." *The Annals of Applied Statistics* 8 (4). Institute of Mathematical Statistics: 2122–49.

Alkema, Leontine, Adrian E Raftery, Patrick Gerland, Samuel J Clark, François Pelletier, Thomas Buettner, and Gerhard K Heilig. 2011. "Probabilistic Projections of the Total Fertility Rate for All Countries." *Demography* 48 (3). Springer: 815–39.

Bijak, Jakub. 2008. "Bayesian Methods in International Migration Forecasting." *International Migration in Europe: Data, Models and Estimates*, 255–88.

Bijak, Jakub, and John Bryant. 2016. "Bayesian Demography 250 Years After Bayes." *Population Studies* 70 (1): 1–19. doi:10.1080/00324728.2015.1122826.

Bryant, John R, and Patrick J Graham. 2013. "Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources." *Bayesian Analysis* 8 (3). International Society for Bayesian Analysis: 591–622.

Camarda, Carlo G, Paul HC Eilers, and Jutta Gampe. 2008. "Modelling General Patterns of Digit Preference." *Statistical Modelling* 8 (4). SAGE Publications Sage India: New Delhi, India: 385–401.

Clark, Samuel J. 2016. "A General Age-Specific Mortality Model with an Example Indexed by Child or Child/Adult Mortality." *arXiv Preprint arXiv:1612.01408*.

Congdon, P, S Shouls, and S Curtis. 1997. "A Multi-Level Perspective on Small-Area Health and Mortality: A Case Study of England and Wales." *Population, Space and Place* 3 (3). Wiley Online Library: 243–63.

Feehan, Dennis M, Mary Mahy, and Matthew J Salganik. 2017. "The Network Survival Method for Estimating Adult Mortality: Evidence from a Survey Experiment in Rwanda." *Demography* 54 (4). Springer: 1503–28.

GBD 2016 Mortality Collaborators (IHME). 2017. "Global, Regional, and National Under-5 Mortality, Adult Mortality, Age-Specific Mortality, and Life Expectancy, 1970-2016: A Systematic Analysis for the Global Burden of Disease Study 2016." *The Lancet* 390 (10100). Elsevier: 1084–1150.

Gelman, Andrew, and Donald B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statist. Sci.* 7 (4). The Institute of Mathematical Statistics: 457–72. doi:10.1214/ss/1177011136.

Girosi, Federico, and Gary King. 2008. *Demographic Forecasting*. Princeton University Press.

He, Chunhua, Li Liu, Yue Chu, Jamie Perin, Li Dai, Xiaohong Li, Lei Miao, et al. 2017. "National and Subnational All-Cause and Cause-Specific Child Mortality in China, 1996-2015: A Systematic Analysis with Implications for the Sustainable Development Goals." *The Lancet Global Health* 5 (2): e186–e197.

IPUMS. 2018. "IPUMS Geo2_KE." https://international.ipums.org/international-action/variables/GEO2_KE#description_section.

Kenya National Bureau of Statistics. 2015. "Kenya Demographic and Health Survey 2014." Rockville, MD, USA. http://dhsprogram.com/pubs/pdf/FR308/FR308.pdf.

Leddy, Robert M. 2017. "Methods for Calculating 5-Year Age Group Population Estimates by Sex for Subnational Areas." Available at: https://www2.census.gov/programs-surveys/international-programs/about/global-mapping/subntl-pop-est-methods-pgs-uscb-dec16.pdf.

Lee, Ronald D., and Lawrence R. Carter. 1992. "Modeling and Forecasting U.S. Mortality." *Journal of the American Statistical Association* 87 (419). [American Statistical Association, Taylor & Francis, Ltd.]: 659–71.

http://www.jstor.org/stable/2290201.

Leslie, Patrick H. 1945. "On the Use of Matrices in Certain Population Mathematics." *Biometrika* 33 (3). JSTOR: 183–212.

Lim, Stephen S, Nancy Fullman, Christopher JL Murray, and Amanda Jayne Mason-Jones. 2016. "Measuring the Health-Related Sustainable Development Goals in 188 Countries:: A Baseline Analysis from the Global Burden of Disease Study 2015." *The Lancet*. York, 1–38.

Linard, Catherine, Marius Gilbert, Robert W Snow, Abdisalan M Noor, and Andrew J Tatem. 2012. "Population Distribution, Settlement Patterns and Accessibility Across Africa in 2010." *PloS One* 7 (2). Public Library of Science: e31743.

Masquelier, Bruno. 2013. "Adult Mortality from Sibling Survival Data: A Reappraisal of Selection Biases." *Demography* 50 (1). Springer: 207–28.

Masquelier, Bruno, Jeffrey W Eaton, Patrick Gerland, François Pelletier, and Kennedy K Mutai. 2017. "Age Patterns and Sex Ratios of Adult Mortality in Countries with High Hiv Prevalence." *AIDS* 31. LWW: S77–S85.

Minnesota Population Center. 2017. "Integrated Public Use Microdata Series, International: Version 6.5 [Dataset]." Available at: https://international.ipums.org/international/.

Plummer, Martyn. 2003. "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna, Austria.

Raftery, Adrian E, Nan Li, Hana Ševčíková, Patrick Gerland, and Gerhard K Heilig. 2012. "Bayesian Probabilistic Population Projections for All Countries." *Proceedings of the National Academy of Sciences* 109 (35). National Acad Sciences: 13915–21.

Schmertmann, Carl P, Suzana M Cavenaghi, Renato M Assunção, and Joseph E Potter. 2013. "Bayes Plus Brass: Estimating Total Fertility for Many Small Areas from Sparse Census Data." *Population Studies* 67 (3). Taylor & Francis: 255–73.

Schmertmann, Carl, and Marcos Roberto Gonzaga. 2018. "Bayesian Estimation of Age-Specific Mortality and Life Expectancy for Small Areas with Defective Vital Records." SocArXiv.

Schmertmann, Carl, Emilio Zagheni, Joshua R Goldstein, and Mikko Myrskylä. 2014. "Bayesian Forecasting of Cohort Fertility." *Journal of the American Statistical Association* 109 (506). Taylor & Francis: 500–513.

Sevcikova, Hana, Adrian E Raftery, and Patrick Gerland. 2017. "Probabilistic Projection of Subnational Total Fertility Rates." *arXiv Preprint arXiv:1701.01787*.

Swanson, David A, and Jeff Tayman. 2012. *Subnational Population Estimates*. Vol. 31. Springer Science & Business Media.

Tatem, Andrew J, Andres J Garcia, Robert W Snow, Abdisalan M Noor, Andrea E Gaughan, Marius Gilbert, and Catherine Linard. 2013. "Millennium Development Health Metrics: Where Do Africa's Children and Women of Childbearing Age Live?" *Population Health Metrics* 11 (1). BioMed Central: 11.

U.S. Census Bureau. 2017. "Subnational Population by Sex, Age, and Geographic Area." Available at: https://www.census.gov/geographies/mapping-files/time-series/demo/international-programs/subnationalpopulation.html.

UN-IGME. 2017. "Levels and Trends in Child Mortality: Report 2017." Available at: http://www.childmortality.org/files_v21/download/IGME%20report%202017%20child%20mortality%20final.pdf.

UNPD. 2017a. "World Population Prospects: The 2017 Edition." Available at: http://esa.un.org/wpp/.

———. 2017b. "World Population Prospects: The 2017 Edition. Methodology of the United Nations Population Estimates and Projections." Available at: https://esa.un.org/unpd/wpp/publications/Files/

WPP2017_Methodology.pdf.

Wachter, Kenneth W. 2014. *Essential Demographic Methods*. Harvard University Press.

Westoff, Charles F, and Anne R Cross. 2006. "The Stall in the Fertility Transition in Kenya." Calverton Maryland ORC Macro MEASURE DHS 2006 May.

Wheldon, Mark C, Adrian E Raftery, Samuel J Clark, and Patrick Gerland. 2013. "Reconstructing Past Populations with Uncertainty from Fragmentary Data." *Journal of the American Statistical Association* 108 (501). Taylor & Francis: 96–110.

———. 2016. "Bayesian Population Reconstruction of Female Populations for Less Developed and More Developed Countries." *Population Studies* 70 (1). Taylor & Francis: 21–37.

World Health Organization (WHO). 2016a. "WHO Methods and Data Sources for Life Tables 1990-2015." Available at: http://www.who.int/healthinfo/statistics/LT_method.pdf.

———. 2016b. *World Health Statistics 2016: Monitoring Health for the Sdgs Sustainable Development Goals*. World Health Organization.

WorldPop. 2018. "Population Movements: Mapping Population Mobility and Connectivity." www.worldpop.org.