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

Monica Alexander*
University of California, Berkeley

Leontine Alkema†
University of Massachusetts, Amherst

Extended abstract submitted to PAA 2018

Abstract

Accurate estimates of subnational-level populations are important for policy formulation and monitoring key 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 incorporates available data on population counts, builds on characteristic mortality schedules and estimates migration patterns to obtain robust population estimates, projections and uncertainty levels. The model is applied to estimate subnational populations in Kenya from 1969-2015, with initial testing showing promising results.

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 global health 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 (WHO (2016), 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 past progress and likely future trajectories.

A population particularly of interest is women of reproductive age (WRA), i.e. those aged 15-49. This subgroup form 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

*monicaalexander@berkeley.edu
†lalkema@umass.edu

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.

In this abstract we present a Bayesian hierarchical model to estimate and project subnational populations of WRA. Building on a cohort component framework, the model uses available data on population counts, estimates migration patterns and incorporates mortality schedules from the national population to estimate and project populations over time. In the remainder of this abstract, we present the model framework, describe some initial results based on the Kenyan population, and outline plans for future work.

2 Model setup

Let $\eta_{r,t}$ be the quantity of interest, i.e. the (true) population of women of reproductive age in each region r at each 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 up the relevant ages (15-49) and cohorts to obtain $\eta_{r,t}$:

$$\eta_{r,t} = \sum_{a;c[t]} \eta_{r,a,c} \tag{1}$$

This cohort perspective allows some demographic structure about mortality and migration trends across age and cohorts to be built into the model.

2.1 Model for true population

We can express the number of women at age a in cohort c as

$$\eta_{r,a,c} = \eta_{r,a-1,c} \cdot \rho_{r,a-1,c} \tag{2}$$

i.e. the number in age group a is the number in the previous age $a-1$ in that cohort times some multiplier. This $\rho_{r,a-1,c}$ term encapsulates mortality and migration in age $a-1$. As such we can express the multiplier as:

$$\rho_{r,a,c} = \phi_{r,a,c} \cdot \psi_{r,a,c} \cdot \delta_{r,a,c} \tag{3}$$

where $\phi_{r,a,c}$ is the mortality component, $\psi_{r,a,c}$ is the migration component, and $\delta_{r,a,c}$ is some error.

2.1.1 Mortality

To estimate the mortality multipliers, $\phi_{r,a,c}$, we want to estimate 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.

$$\phi_{r,a-1,c} = 1 - q_{r,a-1,c} \tag{4}$$

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

Although there is usually not much data available on mortality at the regional level, we can utilize information we know about mortality trends on the national level. We use semi-parametric models that capture shape of national mortality through age and time but allow for differences by region. In particular, we model regional mortality on the logit scale as

$$\text{logit } q_{r,a,c} = \text{logit } \bar{q}_a + \beta_{1,r,c} \cdot Y_{1,a} + \beta_{2,r,c} \cdot Y_{2,a} \tag{5}$$

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.

Principal components are obtained from a decomposition on a matrix which contains a set of standard mortality curves. The mean mortality schedule and the first two principal components for Kenyan national mortality curves from 1950–2020 are shown in Fig. 1 below. These data were obtained from national Kenyan life tables produced by the United Nations Population Division as part of the World Population Prospects 2017 (UNPD, 2017). We used constrained principal components computation to ensure all components were non-negative. This was done to ensure past effects of HIV/AIDs would not be projected into the future.

The mean logit mortality schedule shows a standard age-specific mortality curve. Mortality is relatively high in the younger ages, and increases with age from about age 10. 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 β terms in Eq. 5 are coefficients that need to be estimated in the model. We place a hierarchical structure on the β's and model the differences in adjacent β's to be centered around national differences:

$$\beta_{d,p,c} - \beta_{d,p-1,c} \sim N(B_{d,p,c} - B_{d,p-1,c}, \sigma^2_\beta) \tag{6}$$

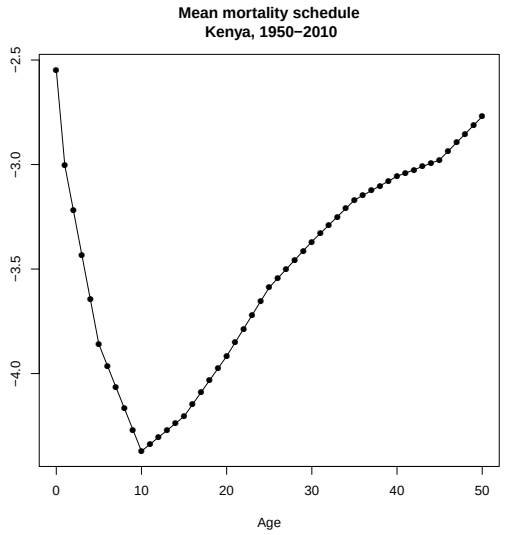
where d refers to the principal component ($d = 1, 2$) and the B's are the principal component coefficients derived from the national mortality schedules. This setup assumes that changes in regional patterns in mortality are distributed around changes in the national mean, with some associated variance, which is estimated in the modeling process.

2.1.2 Migration

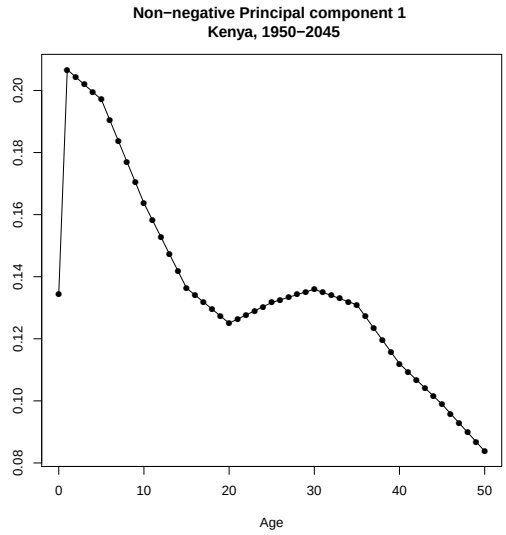
In addition to capturing age-specific mortality, we also want to allow for different migration patterns across age and region. We model the migration term $\psi_{r,a,c}$ as a random effect centered at an age- and region-specific mean:

$$\psi_{r,a,c} \sim N(\mu_{\psi_{r,a}}, \sigma^2_{\psi_{r,a}}) \tag{7}$$

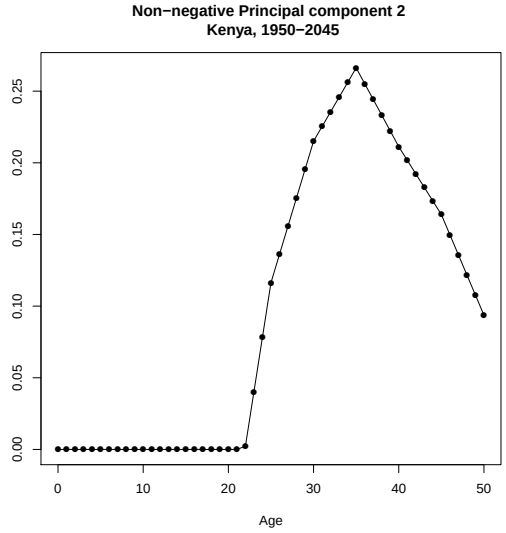
This setup assumes that there are strong age-specific migration patterns that are unique to each region. In future work, we plan to incorporate potential data sources on



(a) Mean schedule



(b) PC1



(c) PC2

Figure 1: First two principal components (demeaned, logit scale), Kenya 1950–2020.

migration flows such as from the Demographic and Health Surveys (DHS) to further inform the structure of the model. For example, in most years of the DHS, survey respondents are asked to report the number of years lived in their current residence, the type of their current residence (e.g.. city, town, countryside) and also the type of previous residence. Information from these questions has the potential to be combined to inform priors in the model about migration inflows and outflows.

2.2 Data model

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

$$y_{r,a,c} \sim N(\eta_{r,a,c}, \sigma_y^2) \tag{8}$$

where σ_y captures measurement and non-measurement error in the data observation.

The raw census data in countries of interest often suffers from a range of data quality issues. In particular, there is often apparent 'age-heaping' on populations on ages ending with '0' or '5'. Before modeling population counts, we smoothed the data across age using splines regression (in future we hope to incorporate this adjustment within the modeling process).

2.3 Model 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 World Population Prospects (WPP) (UNPD, 2017). The WPP models populations of five-year age groups 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. 1970, 1975 ... 2020. These lower and upper bounds are estimated within the model:

$$\begin{aligned} L_{g,y} < & \sum_{a[g],r} \eta_{a,y} & \leq U_{g,y} & (9) \\ \log L_{g,y} & \sim & N(\log 0.9 WPP_{g,y}, 0.1) T(, \log WPP_{g,y}) & (10) \\ \log U_{g,y} & \sim & N(\log 1.1 WPP_{g,y}, 0.1) T(\log WPP_{g,y},) & (11) \end{aligned}$$

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.

2.4 Projection

The model setup defined above allows trends in population counts by age and region to be projected into the future. The β coefficients can be projected forward according to the setup defined in Eq. 6 and these can then be used in combination with the principal components to obtain projections and uncertainty for population counts. In particular, to project the number of women of age a in region r in cohort $c+1$:

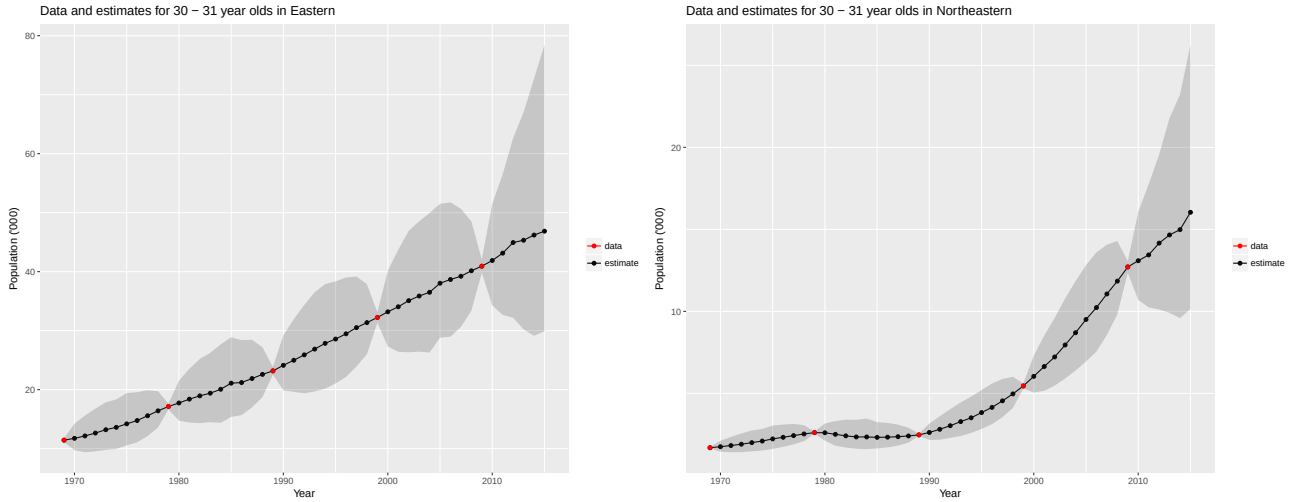
1. Draw values for $\beta_{d,a,c+1}$ based on Eq. 6 and the estimate for σ_β
2. Use these values to calculate the probability of death (based on Eq. 5) and the corresponding mortality multiplier $\phi_{r,a,c+1}$
3. Draw values for $\psi_{r,a,c+1}$ based on Eq. 7 and estimates for $\mu_{\psi_{r,a}}$ and $\sigma_{\psi_{r,a}}$
4. Combine values for $\phi_{r,a,c+1}$ and ψ_{rac+1} to get overall multiplier $\rho_{r,a,c+1}$
5. Calculate value for $\eta_{r,a,c+1}$ based on Eq. 1.

2.5 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.

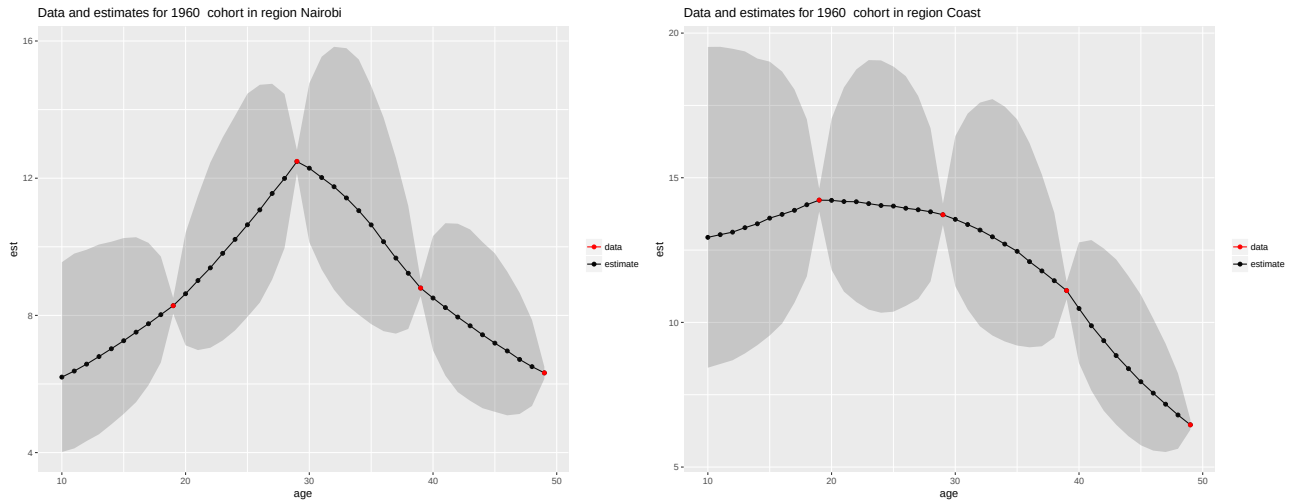
3 Initial results

We tested the model fitting to eight Kenyan Provinces for years 1969–2015. Kenya has dicentennial census data available for years 1969, 1979, 1989, 1999 and 2009. We estimated the female population at each age for 15-49 year olds. To get the total number of women of reproductive age, these estimates can be summed together. For illustrative purposes, Fig. 2 shows estimates of the number of 40 year olds in two provinces over time (panels a) and b)) and the number of women at each age in the 1960 cohort (panels c) and d)). The number of 40 year-olds is increasing in both the Eastern and Northeastern provinces over time, but the rate of increase varies across region. Looking at panel c), it is clear that Nairobi experiences in-migration of the working-age population.



(a) Population of 40-year women over time, Eastern province

(b) Population of 40-year women over time, Northeastern province



(c) Population of 1960 cohort across age, Nairobi

(d) Population of 1960 cohort across age, Coast province

Figure 2: Estimates of 40 year-olds and populations in 1960 cohort by for different provinces, Kenya 1950–2015.

4 Future work

We presented a Bayesian hierarchical model to estimate female populations at the subnational level. The model incorporates available data on population counts, builds on characteristic mortality schedules and estimates migration patterns to obtain robust population estimates and uncertainty levels. Initial testing on Kenya Provinces shows reasonable results.

Future work will focus on incorporating additional information on regional mortality and migration patterns to inform parameters in the model. There is potential to use questions on residence changes in the Demographic and Health Survey (DHS) to get a sense of the direction of migration flows across regions, which could be incorporated as prior information in the model. In addition, there is scope to include information on adult mortality from the DHS to better inform mortality patterns at the regional level.

In addition, we plan to estimate subnational populations at a more granular level; for example in Kenya, this would be producing estimates and projections for around 47 counties. We are also working on parameterizing the extent of age-heaping present in data in order to be able to adjust for data quality issues within the modeling framework.

References

Gelman, A. and D. B. Rubin (1992, 11). Inference from iterative simulation using multiple sequences. *Statist. Sci. 7*(4), 457–472.

He, C., L. Liu, Y. Chu, J. Perin, L. Dai, X. Li, L. Miao, L. Kang, Q. Li, R. Scherpbier, S. Guo, I. Rudan, P. Song, K. Y. Chan, Y. Guo, R. E. Black, Y. Wang, and J. Zhu (2017). National and subnational all-cause and cause-specific child mortality in china, 1996â2015: a systematic analysis with implications for the sustainable development goals. *The Lancet Global Health 5*(2), e186 – e197.

Lee, R. D. and L. R. Carter (1992). Modeling and forecasting u.s. mortality. *Journal of the American Statistical Association 87*(419), 659–671.

Lim, S. S., N. Fullman, C. J. Murray, and A. J. 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*, 1–38.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

United Nations Population Division (UNPD) (2017). World population prospects: The 2017 edition. Available at: `http://esa.un.org/wpp/`.

World Health Organization (WHO) (2016). *World Health Statistics 2016: Monitoring Health for the SDGs Sustainable Development Goals*. World Health Organization.