Demography (2022) 59(5):1713–1737 Published online: 9 September 2022
DOI 10.1215/00703370-10216406

A Bayesian Cohort Component Projection Model to Estimate Women of Reproductive Age at the Subnational Level in Data-Sparse Settings

Monica Alexander and Leontine Alkema

ABSTRACT Accurate estimates of subnational populations are important for policy formulation and monitoring population health indicators. For example, estimates of the number of women of reproductive age are important to understand the population at risk of maternal mortality and unmet need for contraception. However, in many low-income countries, data on population counts and components of population change are limited, and so subnational levels and trends are unclear. We present a Bayesian constrained cohort component model for the estimation and projection of subnational populations. The model builds on a cohort component projection framework, incorporates census data and estimates from the United Nation's World Population Prospects, and uses characteristic mortality schedules to obtain estimates of population counts and the components of population change, including internal migration. The data required as inputs to the model are minimal and available across a wide range of countries, including most low-income countries. The model is applied to estimate and project populations by county in Kenya for 1979–2019 and is validated against the 2019 Kenyan census.

KEYWORDS Population projection • Bayesian methods • Subnational • Cohort component models • Women of reproductive age

Introduction

Reliable estimates of demographic and health indicators at the subnational level are essential for monitoring trends and inequalities over time. As part of monitoring progress toward global health targets such as the Sustainable Development Goals (SDGs), there has been increasing recognition of the substantial differences that can occur across regions within a country (He et al. 2017; Lim et al. 2016; World Health Organization 2016). The analysis of national-level trends is often inadequate, and subnational patterns should be considered to fully understand likely future trajectories. Indeed, estimates and projections of important indicators such as child mortality and contraceptive use are now being published at the subnational level (New et al. 2017; Wakefield et al. 2019).

ELECTRONIC SUPPLEMENTARY MATERIAL The online version of this article (https://doi.org/10.1215/00703370-10216406) contains supplementary material.

To effectively measure health indicators of interest, we need to be able to accurately estimate the size of the population at risk. In order to convert the rate of incidence of a particular demographic or health outcome to the number of people affected by that outcome, we need a good estimate of the denominator of those rates. Hence population counts are essential knowledge for policy planning and resource allocation purposes. However, even something as seemingly simple as the number of people in an area of a certain age is relatively unknown in many countries, particularly low-income countries that do not have well-functioning vital registration systems. And as previously reported outcomes have shown, differences in estimates of the population at risk 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 (UN-IGME 2017; Wang et al. 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.

Data on population counts by age and sex at the subnational level vary substantially by country, and often data availability and quality are the worst in countries where outcomes are also relatively poor. For example, many low-income countries may have only one or two historical censuses available, and very little data available on internal migration or mortality rates at the subnational level. This situation is in stark contrast to many high-income countries, where multiple data sources on population counts, mortality, and migration may exist. These varying data availability contexts present challenges in estimates of both population and the components of population change. In data-rich contexts, the challenge is to reconcile multiple data sources that may be measuring the same outcome. In data-sparse contexts, the challenge is to obtain reasonable estimates without many observations. In both cases, traditional demographic models are often utilized, which often center around a cohort component projection framework and take advantage of the fact that patterns in populations often exhibit strong regularities across age and time. However, these classical methods do not give any indication of uncertainty around the estimates or projections, and incorporating information from different data sources often requires *ad hoc* adjustments to ensure consistency. To overcome these limitations, we propose a method that builds on classical demographic estimation of subnational populations by incorporating these techniques within a probabilistic framework.

In particular, we present a Bayesian constrained cohort component model to estimate subnational populations, focusing on women of reproductive age (WRA)—that is, women 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. The model presented embeds a cohort component projection setup in a Bayesian framework, allowing uncertainty in data and population processes to be taken into account. At a minimum, the model uses data on population and migration counts from censuses, as well as national-level information on mortality and population trends, taken from the United Nations World Population

Prospects (United Nations 2019a). Because data requirements are relatively modest, the methodology is applicable across a wide range of countries and overcomes limitations of previous subnational cohort component methods, which require relatively large amounts of data. Estimates and projections of population by age are produced, as well as estimates of subnational mortality schedules and in- and out-migration flows. Hence, results from the model are helpful in understanding populations at risk of demographic and health outcomes at the subnational level, but also in understanding the drivers of population change and how these may in turn affect trends in indicators of interest.

The following section gives a brief overview of existing methods of subnational population estimation and outlines the contributions of the model proposed here. We then describe the main data sources typically available for subnational population estimation in low-income countries, using counties in Kenya as an example, followed by a detailed description of the proposed methodology. Next, we present results of fitting the model to data in Kenya and validate its out-of-sample projections against the 2019 census. Finally, possible extensions are discussed.

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. First, 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. In some contexts, international migration can also be an important component of demographic change, particularly in regions of armed conflict. Second, when estimating subnational populations, it is important to ensure that 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 so that they match the total. Lastly, data quality and availability are 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 are often lacking or unreliable. This means that it is particularly important at the subnational level to address and quantify uncertainty in population counts and components of change.

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 on the basis of assumptions of constant levels or trends in shares. For example, the U.S. Census Bureau produces 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 population (Leddy 2016).

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):

$$P_t = P_{t-1} + B_{t-1} + I_{t-1} - D_{t-1} - O_{t-1}. \quad (1)$$

This equation is for a total population, but the same accounting equation holds for each age-group separately (where births affect only the first age-group). The cohort component method of population projection (Leslie 1945) takes a baseline population with a certain age structure and projects it forward using age-specific mortality, fertility, and migration rates. Cohort component methods are important because they allow for overall population change to be related to the main components of that change. By estimating population size on the basis of the components of fertility, mortality, and migration, the method allows changes in these components to be taken into account. However, cohort component methods are more data-intensive than extrapolation methods, which is particularly an issue at the subnational level. Especially for developing countries, where well-functioning vital registration systems do not exist, sufficient data on mortality, fertility, and migration are often lacking.

Other methods of subnational estimation involve building regression models that 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, such as births, deaths, voters, or school enrollments (for a detailed review, see Swanson and Tayman 2012). 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.

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. 2014; Raftery et al. 2012; United Nations 2019a), fertility (Alkema et al. 2011), mortality (Alexander and Alkema 2018; Alkema and New 2014; Girosi and King 2008), and migration (Bijak 2008). In terms of estimating the full demographic accounting identity, Wheldon et al. (2013) proposed a method for the reconstruction of past populations. The model embeds the demographic accounting equation within a Bayesian hierarchical framework, using information from available censuses to reconstruct historical populations via a cohort component projection framework. The authors showed that the method works well to estimate populations and quantify uncertainty in a wide range of countries with varying data availability (Wheldon et al. 2016). The method presented in Wheldon et al. (2013) is designed for population reconstruction at the national level, and as such, accounting for internal migration is not an issue. In addition, their method relies on and calibrates to national population estimates produced as part of the UN World Population Prospects.

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 (Alexander et al. 2017; Congdon et al. 1997) and sparse (Schmertmann and Gonzaga 2018). For subnational fertility estimation, Sevcikova et al. (2018) proposed 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) proposed a method that 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, John Bryant and colleagues have shown how the demographic accounting equation can be placed within a Bayesian framework to account for and reconcile different data sources, population counts, and the components of population change (Bryant and Graham 2013; Bryant and Zhang 2018). Bryant and Zhang (2018) showed how the underlying demographic processes can be captured through a process or system model, and different types of uncertainty around data inputs are captured through data models. The focus of Bryant and Graham (2013) was producing subnational population estimates for New Zealand—reconciling and incorporating information about the population from sources such as censuses and school and voting enrollments. The approach that we take in this article is similar to the Bryant et al. approach, in that we model population change with a process model, the components of which are described by system models, and different sources of information are combined through the use of data models. However, whereas Bryant et al. tried to overcome challenges of combining multiple data sources that may be measuring the same outcome, we are trying

to overcome the challenges of estimating subnational populations in contexts where there are extremely limited amounts of data available.

There has been an increasing amount of research using geolocated data and satellite imagery to estimate population sizes and flows in developing countries (Leasure et al. 2020; Wardrop et al. 2018). 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 (Leasure et al. 2020; Linard et al. 2012). While this work contributes to information about subnational populations, the focus and goals of this estimation work are different from our goals in this study. In particular, the goal of much of the WorldPop work is primarily to obtain estimates of total population and population density at a very granular level, rather than to obtain population estimates by age and sex. The 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. More recently, WorldPop's "bottom-up" approach starts with a census or survey for a particular country and uses spatial models to produce spatially granular estimates of populations by age and sex for a single year (Wardrop et al. 2018). However, less attention has been paid to how age distributions across regions change over time. But changes in age distributions are important in understanding broader population change and how this will impact global health indicators of interest. In addition, our approach is grounded in understanding the main components of demographic change—mortality and migration—over time and how they affect population sizes, rather than just estimating the population size as a single outcome.

The methodology proposed here 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) and Bryant and Zhang (2018) (but with a focus on data-sparse situations).

In particular, we introduce a framework to estimate subnational population counts and components of population change that relies on a minimal amount of data that are available for the vast majority of countries worldwide. Observations on subnational population counts and internal migration movements are taken from censuses, but no information on subnational mortality patterns is required. We instead use a mortality model approach based on principal components derived from national mortality schedules. Using principal components for demographic modeling and forecasting first gained popularity after Lee and Carter used the technique as a basis for forecasting U.S. mortality rates (Lee and Carter 1992). More recently, principal components has become increasingly used in demographic modeling, in both fertility and mortality settings (Alexander et al. 2017; Clark 2016; Schmertmann et al. 2014).

While one strength of our approach is being able to estimate components of subnational population change with limited data, another strength of the proposed framework is that it can be readily extended to include other data or estimates. For example,

gridded estimates produced as part of the WorldPop project could conceivably be treated as an additional data input to the model. Conversely, subnational population estimates and uncertainty bounds using our approach may be useful as inputs to help inform WorldPop's estimates.

Data

We aim to estimate female population counts for ages 15–49 per five-year age-group 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: micro-level data from censuses, and national population and mortality estimates from the 2019 World Population Prospects. These data sources are outlined in the following sections.

Overview of Kenyan Geography

In Kenya, the first administrative units are provinces, and the second administrative units are counties. There are eight provinces, including the capital, Nairobi, and 47 counties. The county boundaries have changed over time but have been stable since the 2009 census. We aim to produce estimates of populations of women of reproductive age at the county level based on county boundaries in 2009. Within the model, we also make use of harmonized district boundaries (see the following description), which are slightly larger than counties. There are a total of 35 districts.

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 from 85 different countries. The majority of countries of interest have relatively recent censuses available through IPUMS International. Kenya has decennial censuses available from 1979 to 2009, and we use data from these years to fit the model in the Kenyan context. Micro-level data are not available for the 2019 Kenyan census, but population counts by sex and five-year age-group and county are available through the national statistics office. While we do not use the 2019 data in model fitting, it is used for model evaluation and validation, as detailed in the Model Evaluation section in Results.

In the micro-level IPUMS data, location of residence is reported at the first (province) and second (county) administrative levels, as well as at a harmonized district level. For Kenya, the provinces are stable over time, but before 2009 the county boundaries changed. Hence, we only have data at the county level for Kenya for

2009. However, we can make use of the harmonized districts for data in years prior to 2009. The districts represent slightly larger groups than the 47 Kenyan counties, which are harmonized and temporally stable (IPUMS 2018). In all cases, each 2009 county is completely contained in one unique district.

We use census data to obtain information on two different quantities: observed population counts and observed patterns of in- and out-migration. Female population counts by five-year age-groups for ages 15–49 and subnational administrative region are obtained directly from the IPUMS International microdata. Because these data are samples (most commonly 10%), the microdata are multiplied by the person weights to obtain counts by age and area.[1]

Information on internal migration between counties and districts is also obtained from national censuses. This is based on questions about a migrant's location of residence one year ago. We calculated in- and out-migration counts by age-group, for each region and each census. For the 2009 census, the calculations were based on counties; for all earlier years, the calculations were based on districts.

National Estimates from World Population Prospects

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 2019 (United Nations 2019a). 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.

We use estimates from WPP 2019 in two ways. First, we would like to ensure that the sum of population estimates at the regional level agrees with published estimates at the national level. Hence, the standard national population counts produced by WPP are used as a constraint in the model, subject to uncertainty. The WPP models populations of five-year age-groups every five years from 1950 to 2100.

In addition, national mortality estimates produced by WPP are used as the basis of a mortality model for patterns at the regional level, capturing HIV/AIDS-related patterns of mortality. To generate these patterns, we use estimates from multiple countries in the sub-Saharan region. WPP uses the relationship between infant mortality and the probability of dying between ages 15 and 60, that is, ${}_{45}q_{15}$, to estimate a life table based on Coale–Demeny model life tables (United Nations 2019b). We use estimates of the probability of dying between ages x and $x+5$, or ${}_5q_x$.

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 in online Appendix A.

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

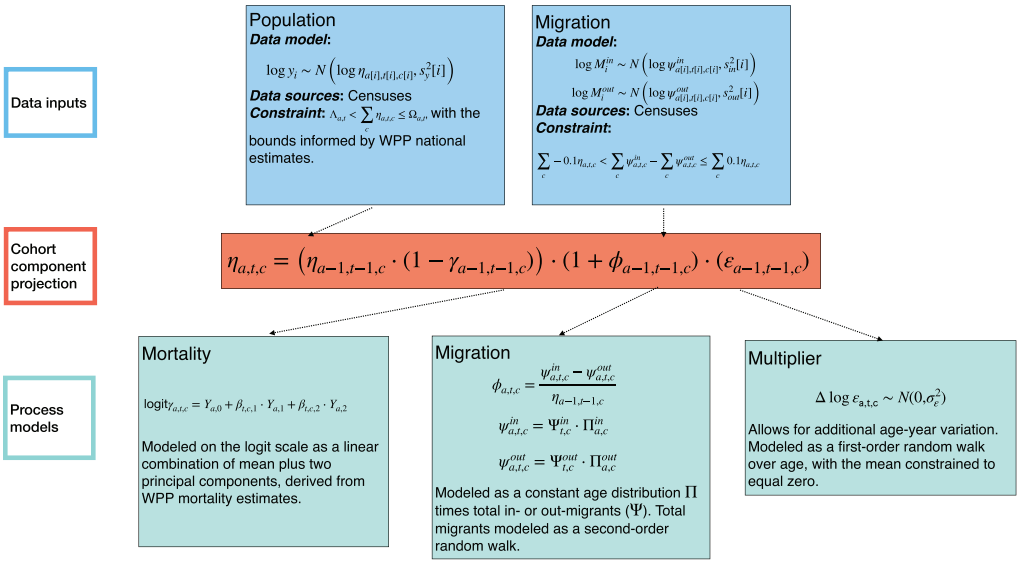

Fig. 1 Diagram showing the main components of the Bayesian cohort component projection model

Model

Overview

In this section, we describe the modeling framework to estimate female populations by five-year age-group and county. The model is outlined for the situation where, as in the Kenyan case, we do not observe county-level information for every census, but we have information on larger, harmonized districts that fully encapsulate the counties. This situation is common for many low-income countries, where geographic boundaries may vary over time but there exist other stably defined boundaries through the microdata on IPUMS.

There are many components and several types of data going into the model at different stages. The overall model framework is summarized visually in Figure 1. We define $\eta_{a,t,c}$ to be the underlying "true" population of women in age-group a, year t, and county c. Our main modeling goal is to obtain estimates and projections of these quantities. The population counts follow a cohort component projection (CCP) model, which assumes that population counts in the current time period are those from the previous period, after taking into account expected changes in mortality and migration. The CCP model also includes an additional age–time multiplier that captures any other variation not already captured by expected changes in mortality or migration. Our setup allows for changes in mortality and migration to be projected forward even if there are no data on these components, and is useful in data-sparse contexts where there is limited information available on the individual components of population change. Note that the goal is to estimate adult populations only and hence the model does not include a fertility component. More details on how estimates in the first age-group are derived are discussed in the following.

As illustrated in Figure 1, the mortality, the migration, and the additional age–time multipliers have additional "process models" (shown in the third row), and data on population counts and migration are related to the underlying process through data models (shown in the top row). The following sections broadly describe each component of the model. The full model specification and details can be found in online Appendix B.

Population Model

The model for population includes the cohort component project model; the data model, which relates observations of population counts from the census to the underlying quantities of interest; and the national-level constraint, which relates the sum of the county-level populations to WPP estimates.

Cohort Component Projection Model

The underlying population $\eta_{a,t,c}$ can be expressed as

$$\eta_{a,t,c} = \left(\eta_{a-1,t-1,c} \cdot \left(1-\gamma_{a-1,t-1,c}\right)\right) \cdot \left(1+\phi_{a-1,t-1,c}\right) \cdot \left(\varepsilon_{a-1,t-1,c}\right), \tag{2}$$

where $\gamma_{a,t,c}$ is the expected conditional probability of death in age-group a, year t, and county c; $\phi_{a,t,c}$ is expected net migration (i.e., in- minus out-migration) as a proportion of population size; and $\varepsilon_{a,t,c}$ is an additional age–year–county multiplier. The multiplier allows for additional changes in age that may not have already been captured by the constrained mortality and migration components.

Note that this is a form of a cohort component projection framework. As mentioned previously, our main modeling goal is to obtain estimates of the $\eta_{a,t,c}$, but we are also interested in estimates of expected mortality ($\gamma_{a,t,c}$) and expected migration ($\phi_{a,t,c}$), and, if nonzero, the log multipliers ($\log \varepsilon_{a,t,c}$).

Data Model

Define y_i to be the ith observed population count. Depending on the year of the census, y_i is observed at either the county c level or district d level. The data model is

$$\log\, y_i \mid \eta_{a,t,c} \;\sim\; \begin{cases} N\left(\log \eta_{a[i],t[i],c[i]}, s_y^2[i]\right) \text{ if } t[i] = 2009, \\ N\left(\log \sum_{c \in d[i]} \left(\eta_{a[i],t[i],c[i]}\right), s_y^2[i]\right) \text{ if } t[i] < 2009, \end{cases} \tag{3}$$

where s_y^2 is the sampling error based on the fact that the microdata in IPUMS are a 10% sample.[2] The second case of this equation dictates that if we have observations prior to 2009, we can relate these only to $\eta_{a,t,c}$s that have been summed to the district level.

[2] Sampling errors were calculated assuming a binomial distribution and using the delta method.

Constraints on National Population

We would like to ensure that the county-level populations $\eta_{a,t,c}$ imply a national-level population that is consistent with previously published estimates in WPP. We would like to use WPP estimates; however, these do not have associated uncertainty published, so we assume the following. We constrain the sum of the county populations by age and year to be within the interval $\left(\Lambda_{a,t}, \Omega_{a,t}\right)$:

$$\Lambda_{a,t} < \sum_c \eta_{a,t,c} \leq \Omega_{a,t}, \tag{4}$$

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:

$$\log \Lambda_{a,t} \sim N\left(\log 0.9\ WPP_{a,t}, 0.1^2\right) T\left(,\log\ WPP_{a,t}\right). \tag{5}$$

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:

$$\log \Omega_{a,t} \sim N\left(\log 1.1\ WPP_{a,t}, 0.1^2\right) T\left(\log\ WPP_{a,t},\right). \tag{6}$$

Note that in this setup, we do not use WPP estimates as "data" to directly inform the sum of the county estimates as in other work (e.g., WorldPop estimates). Instead, we use the WPP estimates to exclude combinations that are extreme as compared to the WPP estimates.

Priors on First Year and Age-group

Because we are interested in estimating and projecting adult populations, the cohort component projection framework does not explicitly take fertility into account. However, the projection framework still requires an initial population of 15-year-olds from which to project populations forward. Additionally, the model requires initial populations in the first year of reconstruction. We place priors on the size of these initial populations using information about the national populations from WPP and the county population proportions from the censuses.

In particular, we use the following priors:

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

$$\log \eta_{a,1,c} \sim N\left(\log\ WPP_{a,1} + \log \text{prop}_{a,1,c}, 0.01^2\right), \tag{8}$$

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-group, year, and county, which was calculated by interpolating census-year proportions and assuming the proportion of a district's population in each county was constant at a level equal to 2009.

Mortality Model

Equation (2) requires estimates of the expected conditional probability of death in each age-group, year, and county. As discussed in the Data section and the online appendix, we do not have reliable information about mortality by age at the county level, and hence we use information about mortality trends at the national level as the basis for a mortality model at the subnational level. A semiparametric model is used to capture the shape of national mortality through age and time, while allowing for differences by county. In particular, we model county mortality on the logit scale as

$$\text{logit}\left(\gamma_{a,t,c}\right) \quad = \quad Y_{a,0} + \beta_{t,c,1} \cdot Y_{a,1} + \beta_{t,c,2} \cdot Y_{a,2}, \tag{9}$$

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

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

$$\beta_{t,c,k} \quad = \quad B_{t,k}^{nat} + \delta_{t,c,k}, \tag{10}$$

$$\delta_{t,c,k} \,|\, \delta_{t-1,c,k}, \sigma_\delta^2 \quad \sim \quad N\left(\delta_{t-1,c,k}, \sigma_\delta^2\right), \tag{11}$$

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

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.

Principal components were obtained from a decomposition of a matrix that contains a set of standard mortality curves. We used national life tables published in the World Population Prospects 2019 for a set of 26 countries in sub-Saharan Africa.[3] These countries were chosen because, like Kenya, they experienced substantial increases in mortality owing to HIV/AIDS.

In particular, let $\mathbf{X}$ be a $N \times G$ matrix of logit mortality rates, where N is the number of years and G is the number of age-groups. In this case, we had $N = 16$ years (estimates every five years from 1950 to 2025) of $G = 7$ age-groups (15–19, 20–24, . . . , 45–49). The singular value decomposition (SVD) of $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{UDV'}, \tag{12}$$

where $\mathbf{U}$ is a $N \times N$ matrix, $\mathbf{D}$ is a $N \times G$ matrix, and $\mathbf{V}$ is a $G \times G$ matrix. The first two columns of $\mathbf{V}$ (the first two right-singular values of $\mathbf{X}$) are $Y_{1:A,1}$ and $Y_{1:A,2}$.

[3] Benin, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Comoros, Democratic Republic of the Congo, Ethiopia, Ghana, Guinea, Guinea-Bissau, Kenya, Madagascar, Mali, Mauritania, Niger, Nigeria, Senegal, Sierra Leone, South Africa, Togo, Uganda, United Republic of Tanzania, Zambia, and Zimbabwe.

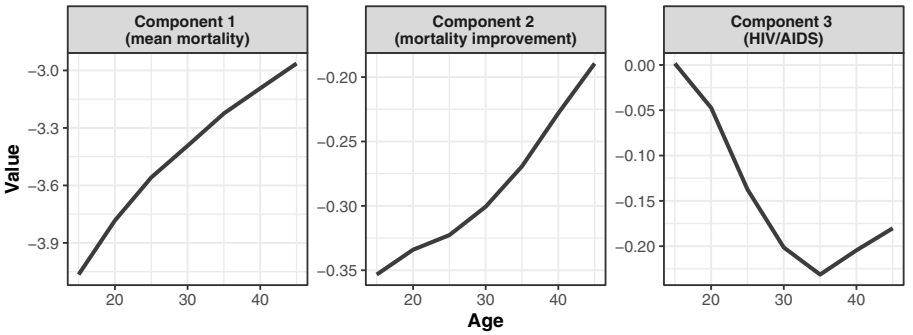

Fig. 2 Mean logit mortality schedule and first two principal components

The mean mortality schedule and the first two principal components for Kenyan national mortality curves between ages 15–49 from 1950 to 2020 are shown in Figure 2. 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 (Sharrow et al. 2014).

Migration Model

The second population change component of Eq. (2) refers to the net-migration rate in a particular age-group, year, and county. Specifically, define the net-migration rate as

$$\phi_{a,t,c} = \frac{\psi^{in}_{a,t,c} - \psi^{out}_{a,t,c}}{\eta_{a-1,t-1,c}}, \tag{13}$$

where $\psi^{in}_{a,t,c}$ is the number of in-migrants and $\psi^{out}_{a,t,c}$ is the number of out-migrants.

For the migration component, we use observed data from the census. In a similar way to the population model, we have a process model, which defines the underlying migration process for the "true" migrant parameters, and a data model, which relates observations from the census to the underlying truth.

Process Model

The model form for the number of in-migrants and out-migrants is informed by patterns observed in the raw census data. In particular, looking at the age distribution of both in- and out-migration (i.e., the proportion of total migrants who are in age-group a) suggests that, while the overall magnitude of migration changes over time, the age patterns in migration are fairly constant (see figures in online Appendix C). This observation is consistent with the large body of migration estimation literature, where strong age patterns of migration motivate both parametric and semiparametric models (e.g., Rogers 1988;

Rogers and Castro 1981; Wisniowski et al. 2015). This observation allows us to simplify the expression for the number of in-migrants and out-migrants, which are modeled as

$$\psi^{in}_{a,t,c} = \Psi^{in}_{t,c} \cdot \Pi^{in}_{a,c}, \tag{14}$$

$$\psi^{out}_{a,t,c} = \Psi^{out}_{t,c} \cdot \Pi^{out}_{a,c}, \tag{15}$$

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. In this way, the age distributions are assumed to be constant over time while the total counts vary.

The total number of in-migrants and out-migrants are estimated for each county–period. We model the total counts as a second-order random walk to impose a certain level of smoothness in the counts over time. As the model captures internal migration flows in and out of each county, it must be the case that the sum of all in-migration flows minus the sum of all out-migration flows equals international net-migration. Lacking sufficient reliable information on international migration for Kenya, we constrain the absolute difference between the sum of all estimated in- and out-migration flows to be less than 10% of the total population for that age-group and period. See online Appendix B for further details.

Data Model

Finally, we relate the observed age-specific in- and out-migration counts in the censuses, denoted M^{in}_i and M^{out}_i, respectively, to the underlying true counts $\psi^{in}_{a,t,c}$ and $\psi^{out}_{a,t,c}$ through the following data model:

$$\log M^{in}_i \mid \psi^{in}_{a,t,c} \sim \begin{cases} N\left(\log \psi^{in}_{a[i],t[i],c[i]}, s^2_{in}[i]\right) \text{ if } t[i] = 2009, \\ N\left(\log \sum_{c \in d[i]} \left(\psi^{in}_{a[i],t[i],c[i]}\right), s^2_{in}[i]\right) \text{ if } t[i] < 2009, \end{cases} \tag{16}$$

$$\log M^{out}_i \mid \psi^{out}_{a,t,c} \sim \begin{cases} N\left(\log \psi^{out}_{a[i],t[i],c[i]}, s^2_{out}[i]\right) \text{ if } t[i] = 2009, \\ N\left(\log \sum_{c \in d[i]} \left(\psi^{out}_{a[i],t[i],c[i]}\right), s^2_{out}[i]\right) \text{ if } t[i] < 2009. \end{cases} \tag{17}$$

In a fashion similar to the data model for population, data observed prior to 2009 can only be related to the migration counts that have been summed to the district level. In addition, the s^2_{in} and s^2_{out} are the sampling errors based on the fact that the microdata in IPUMS are a 10% sample.

Additional Age–Time Multiplier $\varepsilon_{a,t,c}$

In the models for expected mortality and migration discussed in the foregoing, constraints are imposed on the age-specific effects. In particular, the use of the SVD approach to model mortality results in mortality age patterns that are linear combinations of the mean schedule and the components of change (the Y's). Additionally, the migration

model assumes a constant age pattern of migration over time with varying magnitudes of in- and out-migration. We assume these forms in order to greatly reduce the number of parameters that need to be estimated in each model, such that reasonable estimates of mortality and migration rates can still be obtained in data-sparse settings.

To allow for county-specific age and time variation that may not have already been captured by other components, we introduced an additional age–time multiplier $\varepsilon_{a,t,c}$ in the population cohort component model (see Eq. (2)). We model these multipliers on the log scale, and to ensure identifiability we assume that the mean of the sum of the log multipliers over all age-groups is zero. See online Appendix B for more details.

This setup assumes that in general, patterns of population change are well captured by the mortality and migration components, and hence, in the absence of more information, we expect the multiplier terms to be close to zero. Observing many estimated nonzero multiplier terms may suggest that the mortality and migration components need to be reformulated.

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 $\hat{R}$ diagnostic (Gelman et al. 2020) 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.

Data and code are available at https://github.com/MJAlexander/subnational-bayes-ccp.

Results

We illustrate some key results of population counts, mortality, and migration. Additional results are presented in online Appendix E.

Population Estimates and Projections

Population estimates and projections are shown in Figure 3. Part a shows the population of women of reproductive age by province in 1979–2019. The black lines and associated shaded area represent the model estimates and associated 95% credible intervals, respectively. The red dots indicate decennial censuses. Populations of WRA increase in every province, with the largest province being Rift Valley. While Northeastern is the smallest province by population size, the growth rate is relatively rapid, likely owing to the relatively high fertility rates in this province (Kenya National Bureau of Statistics 2015; Westoff and Cross 2006), whereas rapid population increases in Nairobi are driven by in-migration.

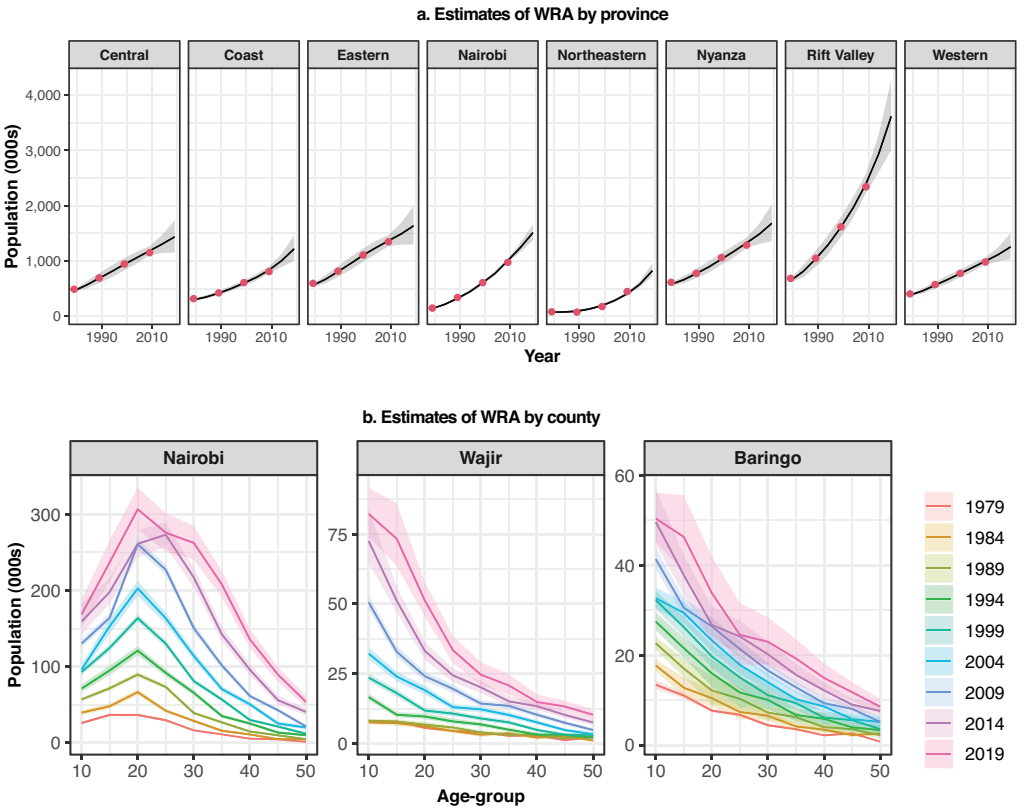

Fig. 3 Population estimates and projections of women aged 15–49 by province and for three counties, by age and year, Kenya, 1979–2019

Part b of Figure 3 illustrates populations over age and time for three counties. Note the different *y*-axis scales for each county. For Nairobi, populations are much larger and the presence of net in-migration far surpasses the effects of mortality, leading to an inverted U-shaped age distribution. For Wajir, a relatively rural county in the northeast, population growth seems rapid over time. For Baringo, populations are relatively small and decline regularly over age owing to mortality.

Mortality

In addition to getting estimates of population counts, we also obtain estimates of the components of population change, namely, mortality and migration. Regarding mortality, there is evidence of variation across the counties. Focusing on three counties as above, mortality profiles are quite different, with Nairobi's estimates being higher than those of the other two counties shown (Figure 4).

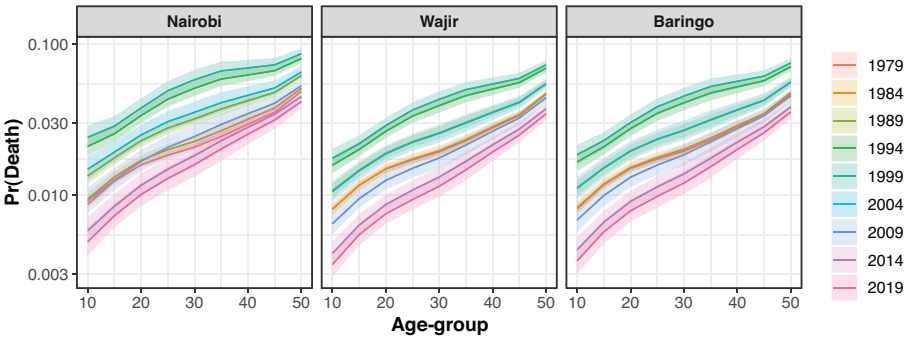

Fig. 4 Estimates of mortality by age and year for three counties

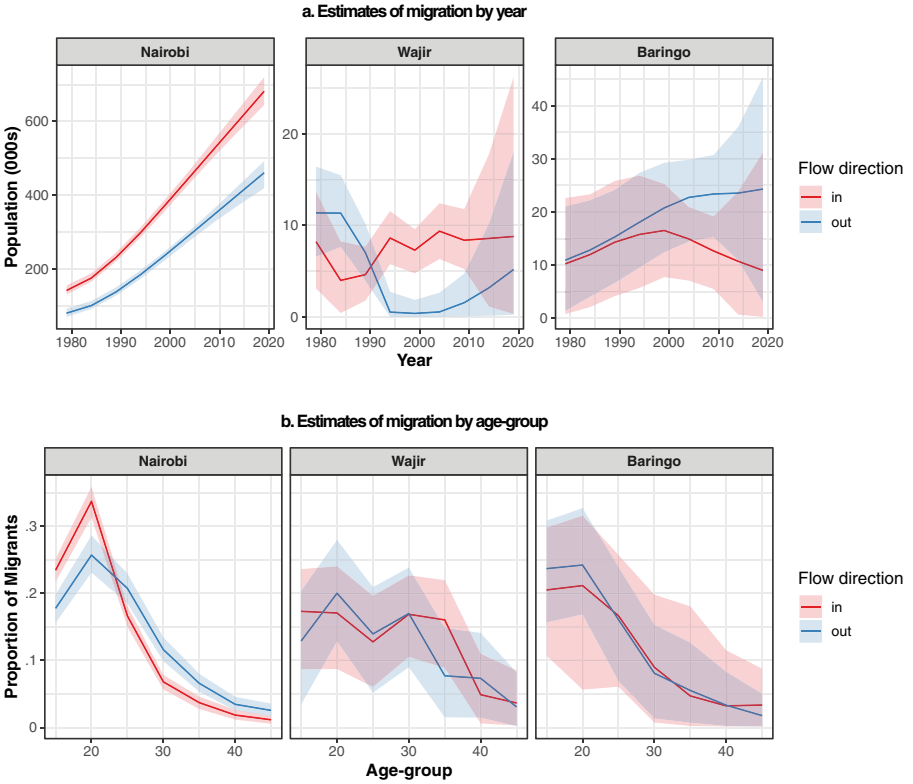

Fig. 5 Estimates of migration components by year and age for three counties

Migration

In addition to mortality, there is substantial variation in patterns in migration across Kenyan counties. Figure 5 shows estimates of all migration components in the three case study counties. For total in-migration and out-migration estimates

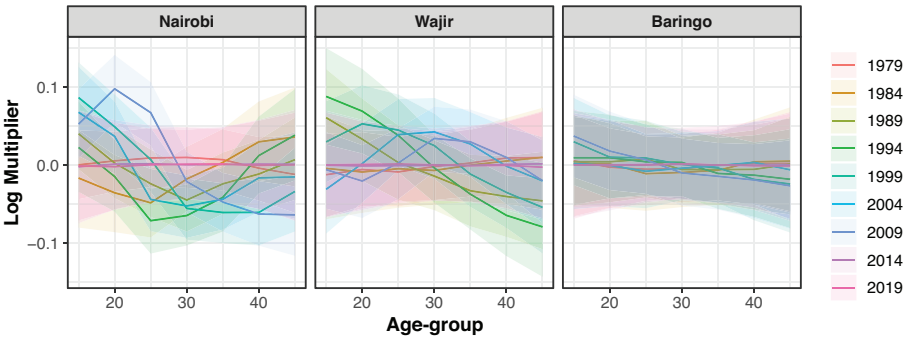

Fig. 6 Estimates of specific age–time multipliers for three counties

(part a), flows into and out of Nairobi are much larger, with net in-migration reaching over 200,000 people per year. Flows into Wajir are much smaller (<10,000 people), and in 2019 Baringo had net out-migration of around 10,000. The estimated age patterns of migration for the three counties are also shown in part b. Some differences exist, with Nairobi's immigrants much more concentrated around ages 20–24.

Age–Time Multiplier

Figure 6 shows the age–time multipliers ε for the three example counties. For Baringo, the multipliers are essentially always zero on the log scale. This observation is true for the majority of counties (see online Appendix E for plots for additional counties), which suggests that most of the patterns over age and time are captured well by the mortality and migration components. For county–years for which multipliers do deviate from zero, estimates are at most around 10% of the total population magnitude, and usually between 0% and 5%. For example, for Nairobi, the estimated multiplier suggests that, after accounting for the expected mortality and migration components, in 2009, we see an additional increase around age 20 (of around 10%) and an additional decrease of almost 10% for the oldest age-group.

Model Evaluation

A national census was conducted in Kenya in 2019. While the micro-level data are not yet publicly available (e.g., via IPUMS), the resulting population counts by age, sex, and county have been published by the Kenya National Bureau of Statistics (Kenya National Bureau of Statistics 2019). We can therefore evaluate the 2019 projections from our model with the actual counts from the 2019 census.

We extracted census population counts by age, sex, and county from a PDF file containing the results following code provided by Alexander (2022). We compared the 2019 projections from the Bayesian cohort component projection model (referred to as Bayes CCP) with these counts and calculated several summary metrics. We define the relative error e_g for a particular group g as

Table 1 Summary of errors in district population sizes by age-group comparing 2019 census counts with two methods, linear interpolation and the Bayesian cohort component projection model (Bayes CPP)

	Mean Error		Median Error		RMSE	
Age-group	Interpolation	Bayes CCP	Interpolation	Bayes CCP	Interpolation	Bayes CCP
15–19	–0.070	–0.058	0.128	0.010	0.013	0.005
20–24	–0.228	–0.081	–0.040	–0.033	0.016	0.005
25–29	–0.266	–0.019	0.018	0.006	0.020	0.005
30–34	–0.146	0.043	0.035	0.047	0.020	0.005
35–39	–0.254	–0.161	–0.008	–0.179	0.035	0.012
40–44	–0.058	–0.074	0.185	–0.048	0.038	0.009
45–49	–0.246	–0.065	0.049	0.011	0.061	0.021
Total Population	–0.101	–0.045	0.119	0.011	0.031	0.010

$$e_g = \frac{y_{g,2019} - \hat{\eta}_{g,2019}}{y_{g,2019}}, \tag{18}$$

where $y_{g,2019}$ refers to the census-based population count for that population and $\hat{\eta}_{g,2019}$ to the model-based projection. A group g can refer to an age–county or age–district group, for example.

On the basis of the errors, we calculated mean, median, and root-mean-squared errors (RMSEs) by age-group and for the total population. We compared these results to the results of a similar linear extrapolation model, where the population in 2019 was estimated by applying the same proportion change seen between the 1999–2009 censuses. Errors were summarized over districts, as estimates by county are not possible with the linear extrapolation method (as we have only one previous set of census observations by county).

Error summaries by age-group and for the total population are shown in Table 1. In general, the Bayesian model projections are within ~1% of the census populations. The magnitudes of the RMSEs for the simple linear interpolation are 3–4 times as high as those of the Bayes CCP. The bias results suggest that the point estimate from the Bayes CCP is often slightly lower than the census observation, whereas linear interpolation substantially overestimates population counts.

We also calculated the coverage of the 95% prediction intervals of the Bayesian CCP model estimates for 2019, compared to the observed 2019 census counts, and the proportion of census counts above and below the prediction intervals. If the model is well calibrated, on average around 90% of the observed census counts should fall within the 90% prediction intervals, and 5% of observations should fall above and below the interval. Table 2 reports coverage by age-group and suggests that in general the coverage of the credible intervals matches expectations. However, in some age-groups, there is a relative bias toward observations falling below the interval rather than above.

We also calculated the probability integral transform (Angus 1994) to assess the consistency between the 2019 projections and observed counts. Results are presented in online Appendix F.

Table 2 Proportion of 2019 census county counts falling within, above, and below the 90% prediction intervals as estimated by the Bayesian CPP model

Age-group	Proportion in Interval	Proportion Above	Proportion Below
15–19	.89	.02	.08
20–24	.89	.00	.09
25–29	.89	.04	.04
30–34	.91	.06	.02
35–39	.87	.01	.09
40–44	.92	.04	.04
45–49	.87	.04	.05

Discussion

In this article, we proposed a Bayesian cohort component projection framework for estimating the population of women of reproductive age when limited amounts of data are available. The model uses information on population and migration counts from censuses, as well as mortality patterns from national schedules, to reconstruct populations from cohorts moving through time. The modeling framework also naturally extends to allow the projection of populations. In addition, the model ensures that the national populations implied by the sum of subnational areas agree with national prepublished UN WPP estimates.

The model was used to estimate and project populations of women of reproductive age for counties in Kenya over the period 1979–2019. Results suggested continued growth of WRA populations in all districts, and accelerated growth in particular in areas such as Nairobi and Northeastern. The mortality component of the modeling framework highlighted the stagnating progress of longevity seen through the 1990s and 2000s, largely due to HIV/AIDS, but also indicated more recent mortality declines. The estimates from the Kenyan example also highlighted substantial differences in internal migration patterns across the nation.

The model requires inputs only 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 alteration. For example, there are currently census microdata available for almost 100 counties on the IPUMS International website. At a minimum, in addition to WPP estimates, data on population counts and in- and out-migration flows by subnational area are ideally required for two census years. These could be obtained via IPUMS as in our case, or using summary aggregate counts if these are available. Such counts at two time points are the minimum requirement, but the more data that are available, the less uncertain estimates are likely to be. The amount of uncertainty and decisions about when it would be appropriate to implement this model are context specific. For example, a country that has relatively stable population growth (with relatively constant or uniformly changing demographic rates) is likely to have less uncertainty around estimates than a country that has experienced fluctuating rates over time. Censuses can be available for any year, as historical, intermediate, and future populations can be reconstructed and projected.

However, note that the longer the projection period (i.e., the more historical the censuses), the larger the projection uncertainty. For countries with substantial international migration, we recommend extending the model to account for this component of population change. For example, if the size of the net-migration flows is known, the constraint on net-migration, which is currently set to be approximately zero, could be updated.

After we ran a series of validation measures, the proposed model outperformed a benchmark model of linear interpolation. In addition to having lower performance than Bayes CCP, with the simple interpolation method it is not possible to easily get estimates by county, because 2009 was the first year that the current geographic boundaries of counties were established. Another advantage of the Bayesian model is that the population estimates also have an associated uncertainty level, and that estimating not only population counts but also mortality and migration rates allows us to better understand the drivers of population change by county.

There are several other advantages and contributions of this modeling framework to the estimation of subnational populations. The model is governed by a cohort component projection model, tracking cohorts as they move through time. This has advantages over more aggregate techniques, such as interpolation and extrapolation, because it allows us to understand trends in overall population as a process governed by separate components that add or remove population. This process takes into account intercensal events, such as trends in HIV/AIDS mortality, and produces estimates and projections with uncertainty.

Second, the modeling framework proposes a parsimonious model for internal net-migration across subnational areas. In cohort component models, migration components are often assumed to be negligible or considered to just be the residual once mortality has been taken into account. Very few data usually exist 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 that separates migration patterns into independent age and time components. The result is an age-specific net-migration model with parameters that are easier to estimate when data are limited.

More broadly, one of the contributions of our proposed framework over existing work in this area is the use of mortality and migration models that have relatively strong functional forms, which allow plausible estimates to be produced even in the absence of good-quality data. Our approach to modeling mortality through the use of characteristic age patterns is inspired by the long demographic tradition of using model life tables when information on mortality is sparse.

While we have illustrated the utility of this approach in data-limited contexts, the framework can naturally be extended to include additional sources of data. For example, if there exist observations of age-specific mortality rates at the subnational level (even at some ages), these data could be used as inputs to the mortality model. If more reliable data on internal migration flows were available, the existing migration process model—which assumes a fixed age schedule with varying magnitude over time—could be reformulated to be more flexible. In general, to be able to handle population projection in a context of low data availability, the model proposed here includes mortality and mortality process models that separate age and time trends into independent effects. Additional specific age–time effects were then captured by the multiplier ε.

If more data are available, the underlying process models could be extended to better understand these age–time effects and how they relate to either mortality or migration.

Another possible extension of this framework is to include other total population estimates, such as those from WorldPop, as additional "data" that could be used to inform estimates. We view this methodology and subnational population estimates produced from it as complementary to estimates produced by efforts such as the WorldPop project. As mentioned earlier, the primary goal of the WorldPop estimates is to produce extremely fine-grained estimates of total population, whereas we are more interested in understanding population patterns by age and sex and the underlying components of population change within larger subnational areas.

This model was developed to estimate and project adult populations only, and in particular, women of reproductive age. The model appears to perform well for this subpopulation. Notably, a limitation of the model is that we do not estimate populations from birth or at older age populations. We do not explicitly model fertility within the cohort component projection framework, but rather use information from censuses and WPP to place plausible prior distributions on the initial age-group of interest (in this case, ages 15–19). A possible extension to the framework proposed here would be the modeling of a larger set of age-groups, starting at birth, and explicitly incorporating fertility rates. As for mortality and migration models, a model for fertility would need to be motivated by the type and amount of data available at the subnational level.

While this model framework has the advantage of explicitly including a component to estimate internal migration flows, a limitation is that we assume there are no biases in the observed migration data. In practice, this may not be true: one can imagine recall biases and "heaping" biases creating underreporting. From exploration of other migration-related data from censuses and Demographic and Health Surveys, it appears these data are the least likely to exhibit such biases. Thus, in the absence of better quality data on migration, we make the assumption that the observations from this census question have no bias.

Another limitation is related to the national constraint in the model. We set the national totals to be between approximately 90% and 110% of the WPP national population estimates. Ideally, we would be able to use uncertainty around the WPP estimates within this model, however, currently this is not available.

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

References

Alexander, M., & Alkema, L. (2018). Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model. *Demographic Research, 38,* 335–372. https://doi.org/10.4054/DemRes.2018.38.15

Alexander, M., Zagheni, E., & Barbieri, M. (2017). A flexible Bayesian model for estimating subnational mortality. *Demography, 54,* 2025–2041.

Alexander, R. (2022). *Telling stories with data: Extracting data from PDFs*. Retrieved from https://tellingstorieswithdata.com/11-clean_and_prepare.html#kenyan-census

Alkema, L., & New, J. R. (2014). Global estimation of child mortality using a Bayesian b-spline bias-reduction model. *Annals of Applied Statistics, 8,* 2122–2149.

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., Pelletier, F., Buettner, T., & Heilig, G. K. (2011). Probabilistic projections of the total fertility rate for all countries. *Demography, 48,* 815–839.

Angus, J. E. (1994). The probability integral transform and related results. *SIAM Review, 36,* 652–654.

Bijak, J. (2008). Bayesian methods in international migration forecasting. In J. Raymer & F. Willekens (Eds.), *International migration in Europe: Data, models and estimates* (pp. 255–282). Chichester, UK: John Wiley & Sons.

Bijak, J., & Bryant, J. (2016). Bayesian demography 250 years after Bayes. *Population Studies, 70,* 1–19.

Bryant, J., & Zhang, J. L. (2018). *Bayesian demographic estimation and forecasting*. Boca Raton, FL: CRC Press.

Bryant, J. R., & Graham, P. J. (2013). Bayesian demographic accounts: Subnational population estimation using multiple data sources. *Bayesian Analysis, 8,* 591–622.

Clark, S. J. (2016). *A general age-specific mortality model with an example indexed by child or child/adult mortality* (ArXiv preprint). https://doi.org/10.48550/arXiv.1612.01408

Congdon, P., Shouls, S., & Curtis, S. (1997). A multi-level perspective on small-area health and mortality: A case study of England and Wales. *Population, Space and Place, 3,* 243–263.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., . . . Modrák, M. (2020). *Bayesian workflow* (ArXiv preprint). https://doi.org/10.48550/arXiv.2011.01808

Girosi, F., & King, G. (2008). *Demographic forecasting*. Princeton, NJ: Princeton University Press.

He, C., Liu, L., Chu, Y., Perin, J., Dai, L., Li, X., . . . Zhu, J. (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. *Lancet Global Health, 5,* e186–e197. https://doi.org/10.1016/S2214-109X(16)30334-5

IPUMS. (2018). *GEO2_KE* [Data set]. Retrieved from 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: The DHS Program, ICF International. Retrieved from http://dhsprogram.com/pubs/pdf/FR308/FR308.pdf

Kenya National Bureau of Statistics. (2019). *2019 Kenya population and housing census volume I: Population by county and sub-county*. Retrieved from https://www.knbs.or.ke/?wpdmpro=2019-kenya-population-and-housing-census-volume-i-population-by-county-and-sub-county

Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V., & Tatem, A. J. (2020). National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty. *Proceedings of the National Academy of Sciences, 117,* 24173–24179.

Leddy, R. M. (2016). *Methods for calculating 5-year age group population estimates by sex for subnational areas* (Population Division Paper, Version 2). Washington, DC: U.S. Census Bureau. Retrieved from https://www2.census.gov/programs-surveys/international-programs/about/global-mapping/subntl-pop-est-methods-pgs-uscb-dec16.pdf

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

Leslie, P. H. (1945). On the use of matrices in certain population mathematics. *Biometrika, 33,* 183–212.

Lim, S. S., Allen, K., Bhutta, Z. A., Dandona, L., Forouzanfar, M. H., Fullman, N., . . . Murray, C. J. L. (2016). Measuring the health-related Sustainable Development Goals in 188 countries: A baseline analysis from the Global Burden of Disease Study 2015. *Lancet, 388,* 1813–1850.

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., & Tatem, A. J. (2012). Population distribution, settlement patterns and accessibility across Africa in 2010. *PloS One, 7,* e31743. https://doi.org/10.1371/journal.pone.0031743

Minnesota Population Center. (2017). *Integrated Public Use Microdata Series, International: Version 6.5* [Data set]. Minneapolis: University of Minnesota. https://doi.org/10.18128/D020.V6.5

New, J. R., Cahill, N., Stover, J., Gupta, Y. P., & Alkema, L. (2017). Levels and trends in contraceptive prevalence, unmet need, and demand for family planning for 29 states and union territories in India: A modelling study using the family planning estimation tool. *Lancet Global Health, 5,* e350–e358. https://doi.org/10.1016/S2214-109X(17)30033-5

Plummer, M. (2003, March). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.* Paper presented at the Third International Workshop on Distributed Statistical Computing, Vienna, Austria. Retrieved from https://www.r-project.org/conferences/DSC-2003/Proceedings/

Raftery, A. E., Alkema, L., & Gerland, P. (2014). Bayesian population projections for the United Nations. *Statistical Science, 29,* 58–68.

Raftery, A. E, Li, N., Sevcikova, H., Gerland, P., & Heilig, G. K. (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences, 109,* 13915–13921.

Rogers, A. (1988). Age patterns of elderly migration: An international comparison. *Demography, 25,* 355–370.

Rogers, A., & Castro, L. J. (1981). *Model migration schedules* (IIASA Research Report, No. RR-81-030). Laxenburg, Austria: International Institute for Applied System Analysis.

Schmertmann, C., Zagheni, E., Goldstein, J. R., & Myrskylä, M. (2014). Bayesian forecasting of cohort fertility. *Journal of the American Statistical Association, 109,* 500–513.

Schmertmann, C. P., Cavenaghi, S. M., Assunção, R. M., & Potter, J. E. (2013). Bayes plus brass: Estimating total fertility for many small areas from sparse census data. *Population Studies, 67,* 255–273.

Schmertmann, C. P., & Gonzaga, M. R. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography, 55,* 1363–1388.

Sevcikova, H., Raftery, A. E., & Gerland, P. (2018). Probabilistic projection of subnational total fertility rates. *Demographic Research, 38,* 1843–1884. https://doi.org/10.4054/DemRes.2018.38.60

Sharrow, D. J., Clark, S. J., & Raftery, A. E. (2014). Modeling age-specific mortality for countries with generalized HIV epidemics. *PloS One, 9,* e96447. https://doi.org/10.1371/journal.pone.0096447

Swanson, D. A., & Tayman, J. (2012). *Springer series on demographic methods and population analysis: Vol. 31. Subnational population estimates*. Dordrecht, the Netherlands: Springer Science+Business Media. Retrieved from https://link.springer.com/book/10.1007/978-90-481-8954-0

UN-IGME. (2017). *Levels and trends in child mortality: Report 2017*. New York, NY: United Nations Children's Fund. Available from http://www.childmortality.org/files_v21/download/IGME%20report%202017%20child%20mortality%20final.pdf

United Nations. (2019a). *World population prospects: 2019* (Report). New York, NY: United Nations, Department of Economic and Social Affairs, Population Division. Available from http://esa.un.org/wpp/

United Nations. (2019b). *World population prospects 2019: Methodology of the United Nations population estimates and projections* (Report). New York, NY: United Nations, Department of Economic and Social Affairs, Population Division. Retrieved from https://esa.un.org/unpd/wpp/publications/Files/WPP2019_Methodology.pdf

U.S. Census Bureau. (2017). *Subnational population by sex, age, and geographic area* [Data set]. Available from https://www.census.gov/geographies/mapping-files/time-series/demo/international-programs/subnationalpopulation.html

Wachter, K. W. (2014). *Essential demographic methods*. Cambridge, MA: Harvard University Press.

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., & Clark, S. J. (2019). Estimating under-five mortality in space and time in a developing world context. *Statistical Methods in Medical Research, 28,* 2614–2634.

Wang, H., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., . . . Murray, C. J. L. (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. *Lancet, 390,* 1084–1150.

Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., . . . Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences, 115,* 3529–3537.

Westoff, C. F., & Cross, A. R. (2006). *The stall in the fertility transition in Kenya* (DHS Analytical Studies No. 9). Calverton, MD: ORC Macro.

Wheldon, M. C., Raftery, A. E., Clark, S. J., & Gerland, P. (2013). Reconstructing past populations with uncertainty from fragmentary data. *Journal of the American Statistical Association, 108,* 96–110.

Wheldon, M. C., Raftery, A. E., Clark, S. J., & Gerland, P. (2016). Bayesian population reconstruction of female populations for less developed and more developed countries. *Population Studies, 70,* 21–37.
Wisniowski, A., Smith, P. W., Bijak, J., Raymer, J., & Forster, J. J. (2015). Bayesian population forecasting: Extending the Lee-Carter method. *Demography, 52,* 1035–1059.
World Health Organization. (2016). *World health statistics 2016: Monitoring health for the SDGs, Sustainable Development Goals*. Geneva, Switzerland: World Health Organization.
WorldPop. (2018). *Population movements: Mapping population mobility and connectivity* [Data set]. Available from www.worldpop.org

Monica Alexander (corresponding author)
monica.alexander@utoronto.ca

Alexander • Departments of Statistical Sciences and Sociology, University of Toronto, Toronto, Ontario, Canada; https://orcid.org/0000-0002-8135-3435

Alkema • Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA, USA; https://orcid.org/0000-0001-8806-5957