Exploration of methods to estimate stateless populations

Monica Alexander and Michael Chong

1 Introduction

This report summarizes three modeling approaches that were explored with the aim of estimating stateless populations in different data contexts. Two approaches are illustrated using different case studies, and a more general approach to estimating adjustments to census data is also discussed.

Specifically, firstly we illustrate how a probabilistic demographic projection framework can be used to reconstruct past populations of Shona in Kenya. While this particular case study involves reconstruction of historical populations, we discuss how this model framework could be used in many different data availability situations, including to forward project future populations, and incorporating multiple, and potentially incomplete data sources.

Secondly, we discuss a conceptual framework to adjust information about statelessness and unknown citizenship reported in censuses, which could be more broadly applied to many countries. The main idea is that if we have countries that have both census data and reasonable quality stateless counts from another source, we can calculate a set of adjustment ratios (by age, and potentially over time). These adjustment ratios can then be incorporated into a statistical framework to adjust census data from other countries where good quality data do not exist. However, after completing an extensive review of census and other data sources, we found there was limited situations where this could be applied. In our view this approach could be potentially applied if a more extensive search for censuses was undertaken, or if this approach was combined with demographic projection.

Lastly, we outline the case study of estimating stateless populations in the United States. We draw on data from the American Community Survey to estimate trends in populations at risk of statelessness, then discuss avenues to potentially convert those at-risk to actual stateless populations. Again this is a conceptual discussion with suggestions made for further data to be incorporated.

2 Demographic projection case study: estimating the Shona population in Kenya over time

In general, available data and estimates of stateless populations may not be complete or up-to-date. In such cases, we may want to construct a current estimate of a population of interest by using a past estimate and "projecting" the population forward in time by simulating births and deaths, and incorporating information about migration where applicable.

The following sections demonstrate the use of this method on the Shona population in Kenya. Since the Shona population was recently enumerated in a survey, we can test the projection method to reconstruct the population in the past. In addition to setting up the machinery to perform the estimation in other contexts, there are several goals for completing this exercise. Estimates of the past and current population are generated under different model conditions to understand how the population estimate behaves under different model assumptions. Primarily, we test different ways of incorporating information in the model that may suit different data contexts.

2.1 Data

This method requires an age-stratified estimate of the population at some point in time, and demographic rates over the desired projection time period. We use the data from the 2019 Kenya Shona survey as a current estimate of the Shona population, and we use the national estimates of fertility and mortality from the United Nation's 2019 World Population Prospects (WPP) for Kenya as the basis for the demographic rates.

The 2019 survey contains responses from 464 households, and contains information on household and individual characteristics. Importantly, demographic information was recorded, including the dates of birth and sex of each individual in the household. This is then aggregated to produce age-specific population counts.

National fertility and survival estimates produced as part of WPP are used. National estimates for Kenya span the period 1950-2019 and contain age-specific rates suitable for an age-structured model.

2.2 Methods

The cohort component projection method reconstructs or forecasts age-specific population counts by using known fertility, survival, and, if applicable, migration information. Generally this involves starting with an initial population estimate and simulating births, deaths, and migrations according to some assumed rates over some period of interest.

One way to perform cohort component projections is to use a Leslie matrix approach. In this approach, age-specific population counts are held in a vector, and demographic rates are organized into a structured matrix (called a Leslie matrix) such that multiplication with the population vector "evolves" the population forward a fixed length of time. Iteratively applying the Leslie matrix thereby approximates the population at discrete points in time.

For age groups $a = 1, \ldots, A$, and periods $t = 1, \ldots, T$ let $s_{a,t}$ denote the survival rate of age group a over period t, which is the proportion of individuals in age group a that survive to the next age group $a + 1$. Similarly, let $f_{a,t}$ denote the age-specific fertility rates of women in age group a over period t. Fertility is assumed to be positive only for age groups $[15, 20), \ldots, [45, 50)$, which we collectively denote as A_f. For other age groups $a \notin A_f$, fertility is assumed to be zero. The vector $\vec{n}_t = (n_{1,t}, \ldots, n_{A,t})$ denotes the population vector at time t where $n_{a,t}$ denotes the size of the population in age group a at the beginning of period t. If the Leslie matrix containing rates over this period is denoted L_t, then the population at the beginning of the next period is calculated

$$\vec{n}_{t+1} = L_t \cdot \vec{n}_t$$

assuming there is no migration. Following the notation of Wheldon et al. (2013), the Leslie matrix L_t is constructed

$$L_t = \begin{bmatrix} \tilde{f}_{0,t} & \tilde{f}_{5,t} & \cdots & \tilde{f}_{A-5,t} & \tilde{f}_{A,t} \\ s_{5,t} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & s_{A,t} & s_{A+5,t} \end{bmatrix}$$

The first row describes population changes due to fertility, where each term

$$\tilde{f}_{a,t} = 5 \cdot s_{0,t} \cdot p_F \cdot \frac{f_{a,t} + f_{a+5,t} \cdot s_{a+5,t}}{2}$$

expresses the number of female children per woman who survive to the next age group. Here, p_F is the proportion of female births, and the factor of 5 adjusts the single year rates

to match the 5-year projection intervals.

The off-diagonal entries contain survival rates which ages the population while accounting for mortality.

If there is migration, then the population at the beginning of the next period is instead calculated

$$\vec{n}_{t+1} = L_t \cdot \left(\vec{n}_t + \frac{\vec{m}_t}{2} \right) + \frac{\vec{m}_t}{2}.$$

where m_t denotes the net number of migrants entering the population at each age group over the period. To improve the discrete-time approximation, this setup assumes that half of the migration occurs at the beginning of the period and the other half occurs at the end of the period such that half of the migrants experience fertility and mortality. In our particular example for the Shona population however, we assume there is no migration.

The demographic rates in a Leslie matrix are usually fixed based on assumptions. Wheldon et al. (2013) introduced a method to perform these projections probabilistically using possibly incomplete data. In their framework, the true underlying population counts and demographic rates are treated as unknown parameters. Intuitively, since different fertility and mortality rates lead to different population sizes and age structures, intermittent population observations such as censuses and surveys can inform what demographic rates likely led to the observed population.

Estimation is performed in a Bayesian setting. Following the convention of Wheldon et al. (2013), we use an asterisk $(*)$ to distinguish observed values from the true values. For example, let $n_{a,t}$ denote the true population counts, and let $n^*_{a,t}$ denote observed population counts where available. The population observations are modeled

$$n^*_{a,t} \sim \text{Normal}(n_{a,t},\, \sigma^2)$$

The true underlying population counts are governed by the Leslie matrix process. Assuming there is no migration, for $t = 1, \ldots, T$ the population vectors are generated as

$$\vec{n}_{t+1} = L_t \cdot \vec{n}_t,$$

and priors are placed on the rates contained in L_t centered at the national rates

$$\log f_{a,t} \sim \text{Normal}(\log f^*_{a,t}, \sigma^2_f) \;\; \text{for } a \in A_f \text{logit}(s_{a,t}) \sim \text{Normal}(\text{logit}(s^*_{a,t}), \sigma^2_s).$$

Population counts are in this way determined iteratively from the initial population count.

However, in this particular application we do not have a reliable estimate of the initial population. We test several setups for the initial population.

2.2.1 Age structure: uniform counts

In the first setup, the population is assumed to have a uniform age distribution, where the count for each age group $a = 1, \ldots, A$ is modeled

$$\log(n_{a,0}) \sim \text{Normal}(\log(100), 0.5).$$

This is the most uninformative model set-up.

2.2.2 Age structure: Dirichlet parameterization

We then try a more realistic setup whereby the initial age distribution is assumed to be similar to that of Zimbabwe in the same period. Zimbabwe was chosen as it was the origin for the majority of the Shona population.

The population age proportions $\vec{p_0} = (p_{1,0}, \ldots, p_{A,0})$ in this case follow a Dirichlet distribution

$$\vec{p_0} \sim \text{Dirichlet}(d \cdot \vec{\alpha}),$$

where $\vec{\alpha} = (\alpha_1, \ldots, \alpha_A)$ is a simplex describing age proportions in the Zimbabwean population, and d is a constant controlling the "concentration" of the age distribution around proportions $\vec{\alpha}$.

Unfortunately, this set-up was difficult to estimate.[1]

We therefore implement this model using the Gamma parameterization of the Dirichlet distribution, which partially addresses the numerical issues. If $\vec{\alpha}' = d \cdot \vec{\alpha} = (\alpha'_1, \ldots, \alpha'_A)$, then an equivalent parameterization of $\vec{p_0}$ is

$$\vec{p_0} = \frac{(\gamma_1, \ldots, \gamma_A)}{\sum_{i=1}^{A} \gamma_i},$$

where $\gamma_i \sim \text{Gamma}(\alpha'_i, 1)$.

[1] Hamiltonian Monte Carlo has trouble sampling directly from this Dirichlet distribution because some of the older age group proportions are very small, creating unreliable divergent transitions.

The total population count is controlled separately. Let $N_0 = \sum_{a=1}^{A} n_{a,0}$,

$$\log(N_0) \sim \text{Normal}(7, 0.7)$$

and the initial age-structured population is obtained by multiplying the proportions by the total population

$$\vec{n}_0 = N_0 \cdot \vec{p}_0.$$

2.2.3 Age structure: Normal log-ratio parameterization

We also tried incorporating the age structure using a Normal log-ratio model. Given A age groups, if $(\alpha_1, \ldots, \alpha_A)$ denote the known Zimbabwean initial age proportions, then define $\vec{r}$ as the vector of log ratios of these proportions, using the middle age category as the baseline:

$$(r_1, \ldots, r_A) = \left(\log \frac{\alpha_1}{\alpha_{A/2}}, \ldots, \log \frac{\alpha_A}{\alpha_{A/2}} \right).$$

The unknown proportions of the Shona population are modeled as

$$\vec{p}_0 = \text{softmax}(\vec{\rho}) = \frac{(\exp(\rho_1), \ldots, \exp(\rho_A))}{\sum_{a=1}^{A} \exp(\rho_a)}$$

where each ρ_a is centered at the observed ratio r,

$$\rho_a \sim \text{Normal}(r_a, 1).$$

The total population count N_0 follows a separate distribution,

$$\log(N_0) \sim \text{Normal}(7, 0.7),$$

and the age-structured proportion is similarly obtained by multiplying the proportions by the total population,

$$\vec{n}_0 = N_0 \cdot \vec{p}_0$$

2.2.4 Data incorporation: counts and proportions

We also explored alternative ways of incorporating the survey observation. Recall that in the above setup, we model the age-specific counts as independently arising from their

corresponding true latent count,

$$n^*_{a,t} \sim \text{Normal}(n_{a,t},\, \sigma_n^2).$$

In some cases however, it may make sense to decouple the total population count from the age details in the data. This may be a sensible option when the total count can be considered reliable but the age data are either (partially or totally) unavailable or considered unreliable.

The idea and setup are similar to that of using age proportion information in the initial population. Given some observation $\vec{n}^*_t = (n^*_{1,t}, \ldots, n^*_{A,t})$, we calculate the observed total $N^*_t = \sum_{a=1}^{A} n^*_{a,t}$ and the observed proportions $\vec{p}^*_t = \vec{n}^*_t / N^*_t$.

We test two options for modeling the total population. In the first option, the observed population total is modeled as

$$\log N^*_t \sim \text{Normal}(\log N_t, \sigma_N^2),$$

where N_t refers to the latent true total population count. We also test modeling on the natural scale such that the observed population total is distributed

$$N^*_t \sim \text{Normal}(N_t, \sigma_N^2).$$

Next, the true latent proportions are modeled with the Normal log ratios parameterization. Let $(p_{1,t}, \ldots, p_{A,t})$ denote the true latent proportions, and $(\rho_{1,t}, \ldots, \rho_{A,t})$ denote the log-ratios relative to the middle age group,

$$(\rho_{1,t}, \ldots, \rho_{A,t}) = \left(\log \frac{p_{1,t}}{p^*_{A/2,t}}, \ldots, \log \frac{p^*_{A,t}}{p^*_{A/2,t}} \right).$$

These log-ratios are centered at the observed values,

$$r_a \sim \text{Normal}(\rho_{a,t}, 1),$$

where $(r_1, \ldots, r_A)$ similarly denotes the observed log-ratios relative to the middle age group,

$$(r_1, \ldots, r_A) = \left(\log \frac{p^*_1}{p^*_{A/2}}, \ldots, \log \frac{p^*_A}{p^*_{A/2}} \right).$$

This setup is tested using the Zimbabwean age structure model similarly with the Normal log-ratio parameterization described in Section 2.2.3.

2.3 Results

2.3.1 Age structure

2.3.1.1 Uniform As a point of comparison, we use a uniform age structure on the initial population. This was useful as a 'baseline' model to see how additional information on the age structure of the population changed estimates. We test the model first with no survey data (i.e. informed only by weak priors), then incorporate the survey data to constrain the population. The population reconstructions under these settings are shown in Figure 1.

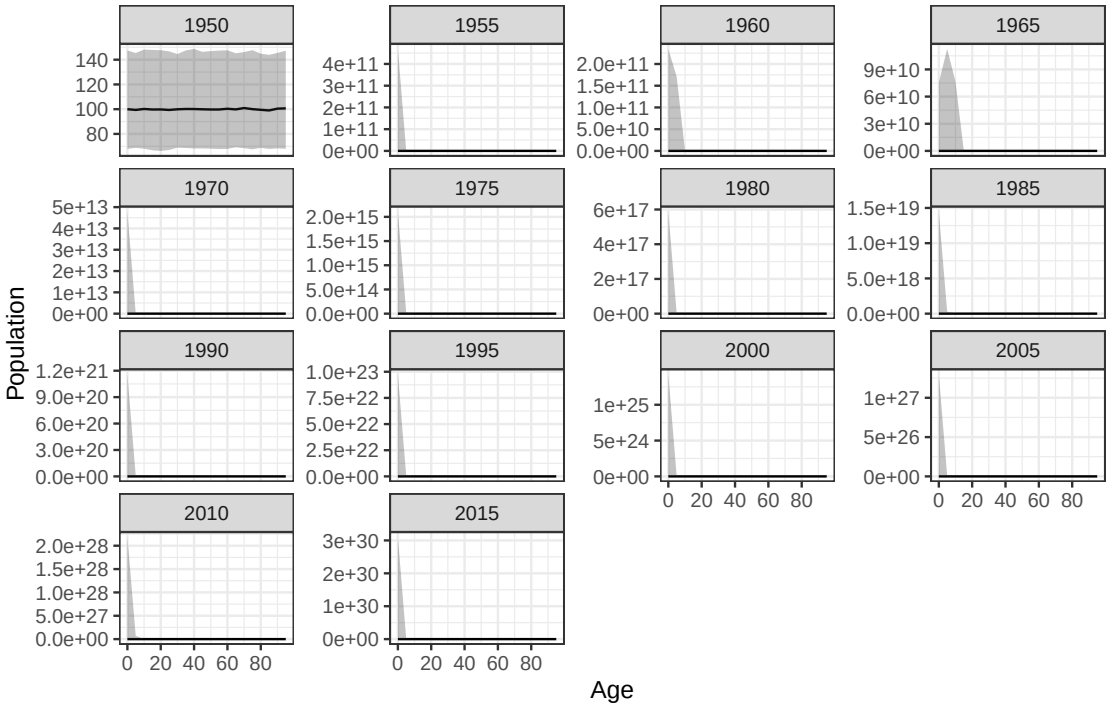
Without the survey observation, the population sizes are poorly constrained and grow exponentially given the survival and fertility rates, which are centered at the national WPP rates for Kenya. Incorporating the survey observation drastically changes the estimated population sizes and decreases the uncertainty around the estimate. The data also have an effect on informing the original age structure.

2.3.1.2 Informed initial age structure: Dirichlet parameterization Here we will present results from a model that uses the Zimbabwean age structure to inform initial age population proportions. We again fit the model in the absence of the 2019 survey data and after incorporating the survey data. The population reconstructions are shown in Figure 2.

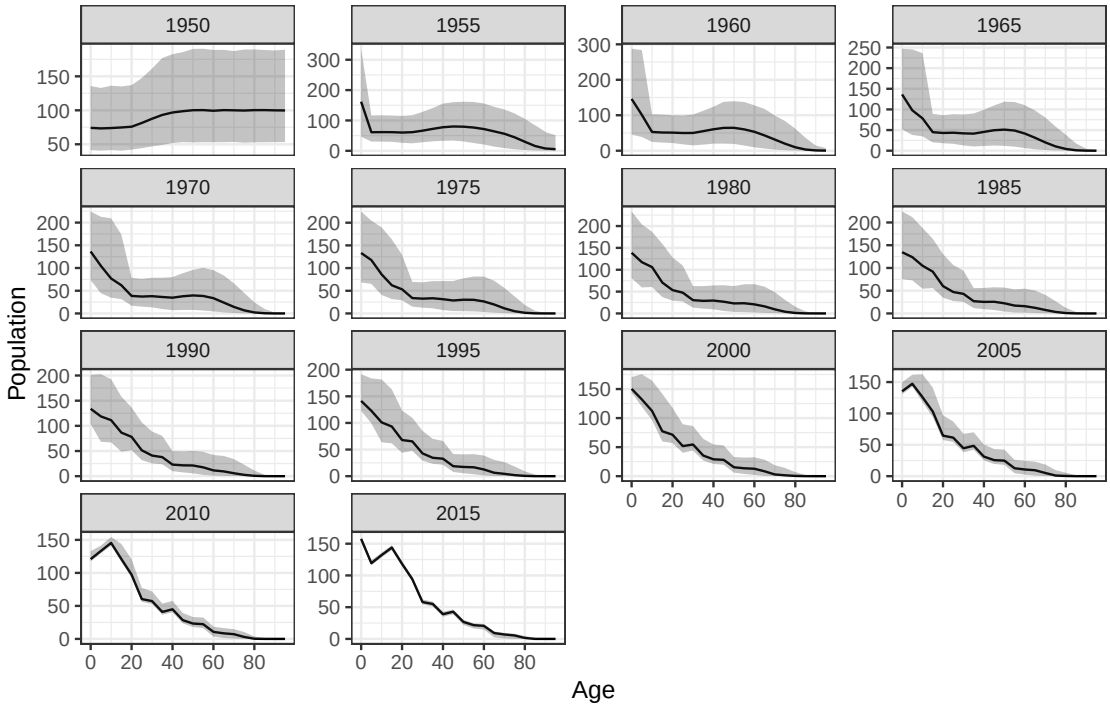
Similar to the first case, the addition of the survey observation is important in constraining the population sizes. The initial population size and the associated uncertainty is constrained after incorporating the observation. However, the initial age distribution is unchanged, as shown in Figure 3.

2.3.1.3 Informed initial age structure: log-Normal parameterization Figure 4 shows the population reconstruction using the log-Normal parameterization for the initial age proportions. Compared to the Dirichlet model in Figure 2, the uncertainty intervals surrounding the age curves are wider in the earlier periods, but are similarly narrow in more recent periods.

Using the log-Normal parameterization, the shape of the initial age distribution is allowed more flexibility. Figure 5 shows the estimated age proportions before and after incorporating the 2019 survey observation with the Zimbabwean proportions as a reference. Despite the large uncertainty intervals surrounding the posterior age structure, there is a noticeable difference in the shape, with a new peak around ages 35-50. This perhaps suggests that the

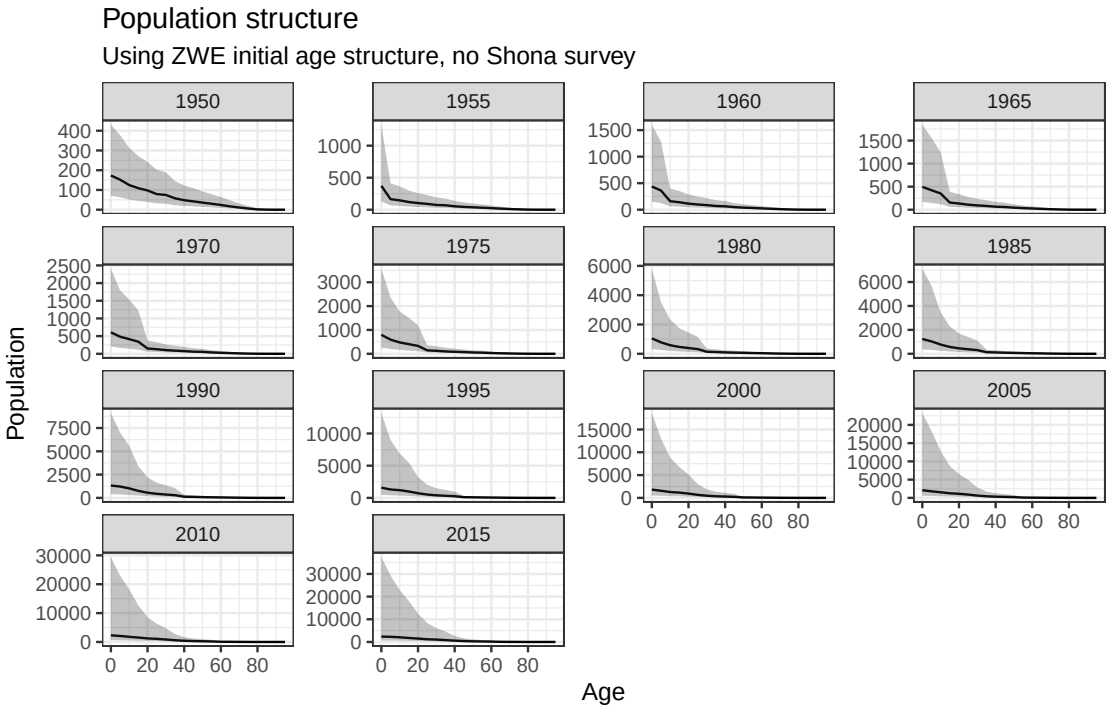


(a) without incorporating survey data

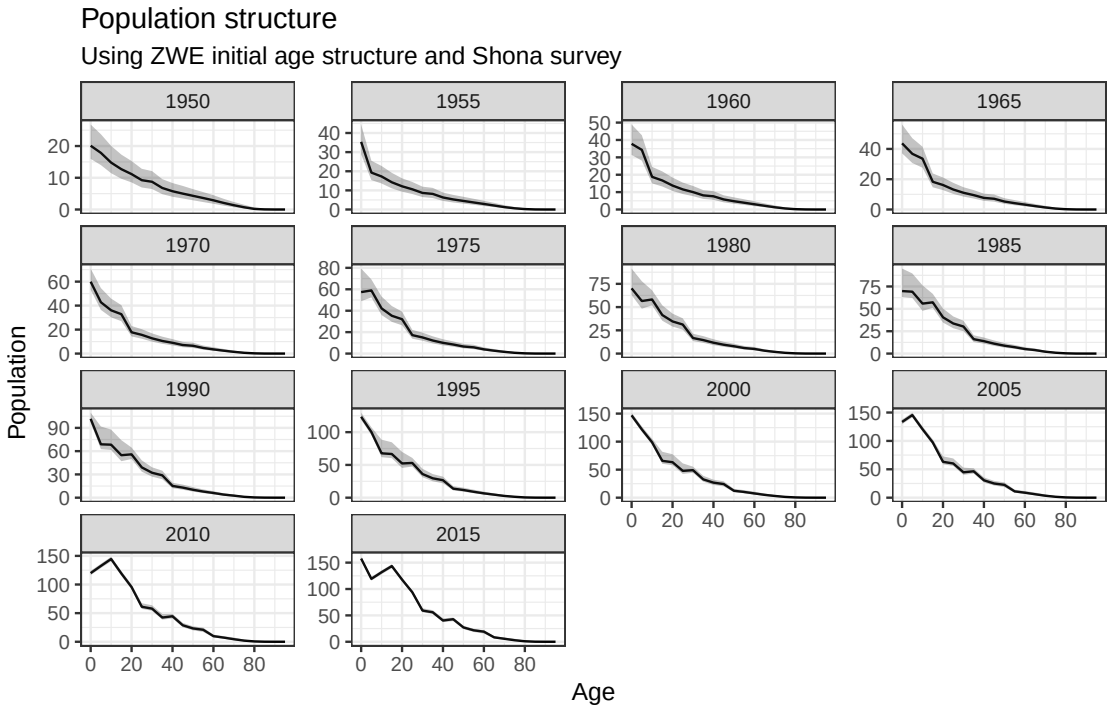


(b) with survey data incorporated

Figure 1: Kenya Shona population reconstruction using uniform initial age structure



(a) without incorporating survey data



(b) with survey data incorporated

Figure 2: Kenya Shona population reconstruction using ZWE initial age proportions

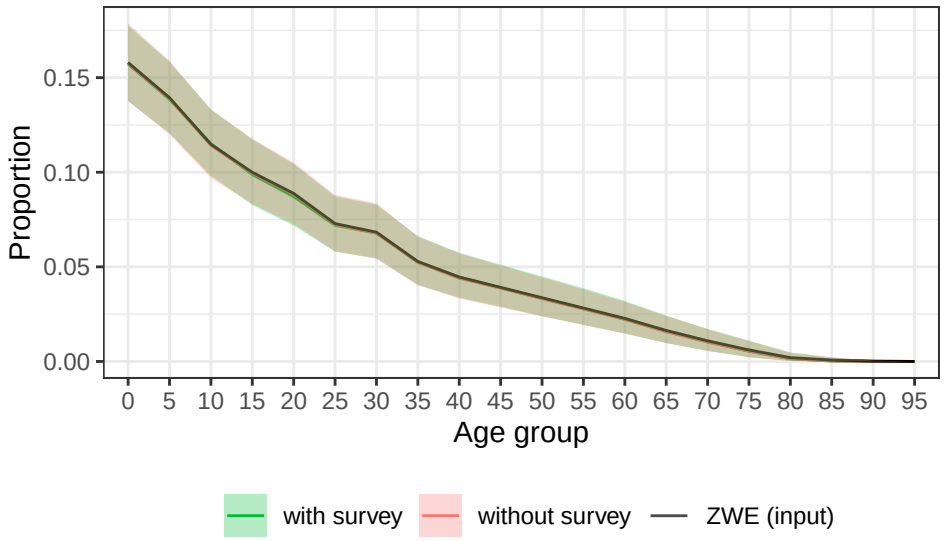


Figure 3: Estimated initial (1950) population age distributions using the Dirichlet model of the Kenya Shona population. Median estimates with and without incorporating the 2019 survey observation are shown in blue and red respectively. Shaded bands represent an 80% UI. The Zimbabwean proportions used to center the priors are shown in black.

Dirichlet model was too restrictive with respect to the age distribution, however the actual point estimates are similar.

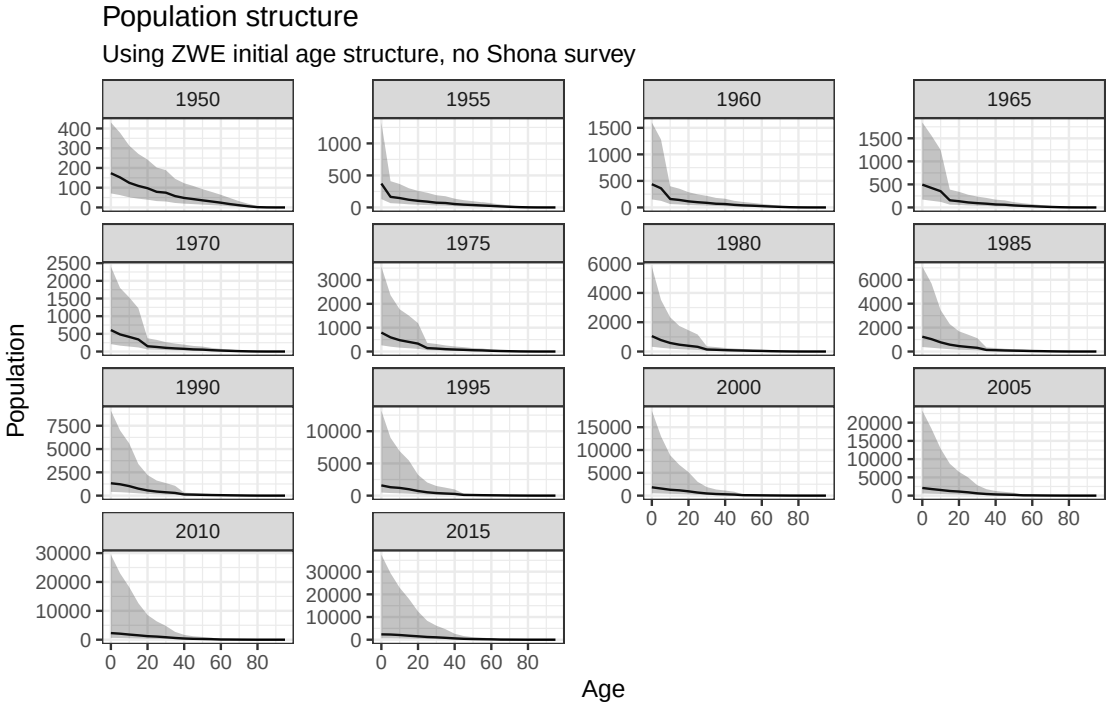
2.3.1.4 Comparisons: fertility and survival rates and population

If the estimated initial populations exhibit different age structure, then we would also expect other parameters in the model to vary in order to compensate such that the final population is consistent with the survey observation. Figures 6, 7, and 8 compare the fertility rates, survival rates, and total populations under each model.

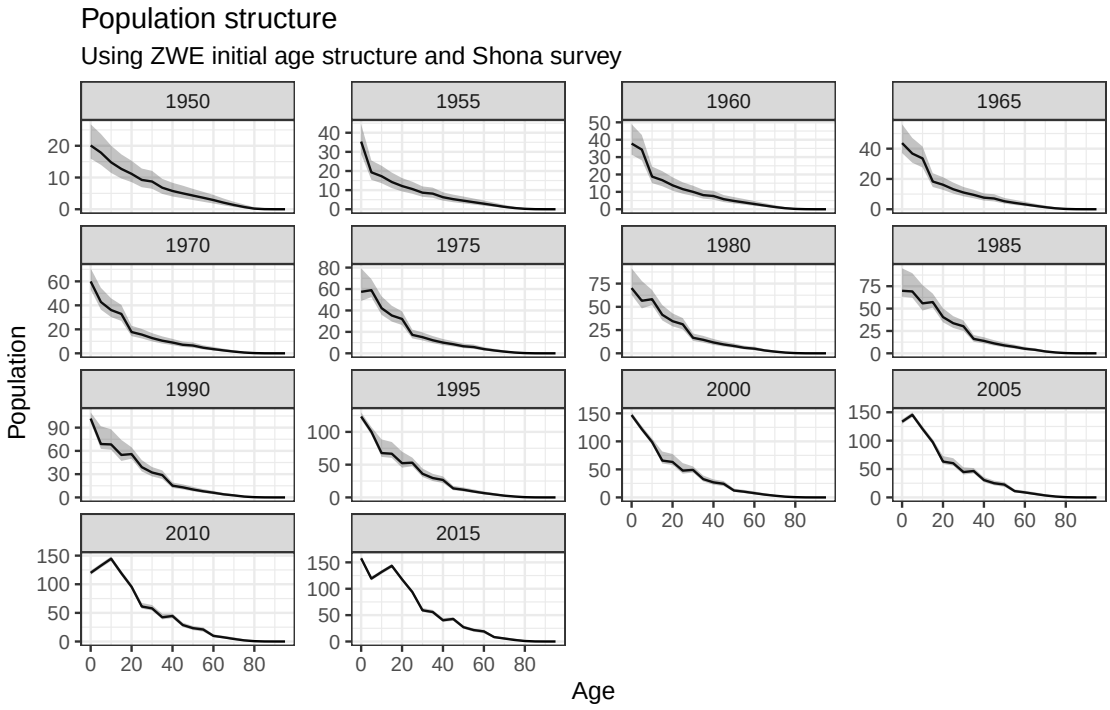
Fertility rates in Figure 6 are generally similar between all model variations, with the exception of some periods (e.g. 1985 and 2000) where the models with informed age structure differ slightly from the uniform age structure after including the 2019 survey observation.

On the other hand the survival rates in Figure 7 show an unusually large increase in infant mortality in 1955-1965 in the uniform age structure model. This might indicate that this initial population is not consistent with the other inputs, which, loosely speaking, the model tries to "correct for" by adjusting the survival rate.

The total populations over time are also quite different. Figure 8 shows the population estimated by the uniform model is much higher than the informed age models, which emphasizes the sensitivity of inferences to the model assumptions.



(a) without incorporating survey data



(b) with survey data incorporated

Figure 4: Kenya Shona population reconstruction using ZWE initial age proportions and log-Normal parameterization

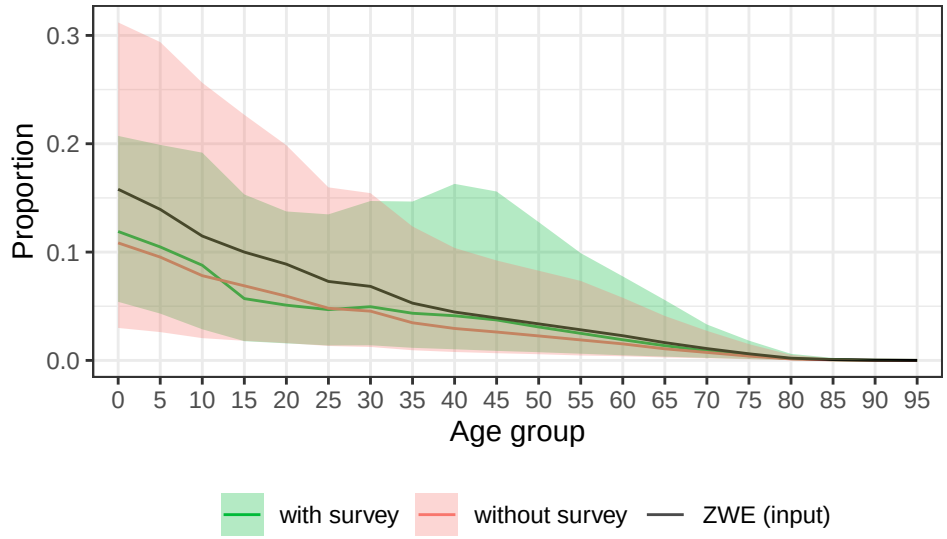


Figure 5: Estimated initial (1950) population age distributions using the log-Normal model of the Kenya Shona population. Median estimates with and without incorporating the 2019 survey observation are shown in blue and red respectively. Shaded bands represent an 80% UI. The Zimbabwean proportions used to center the priors are shown in black.

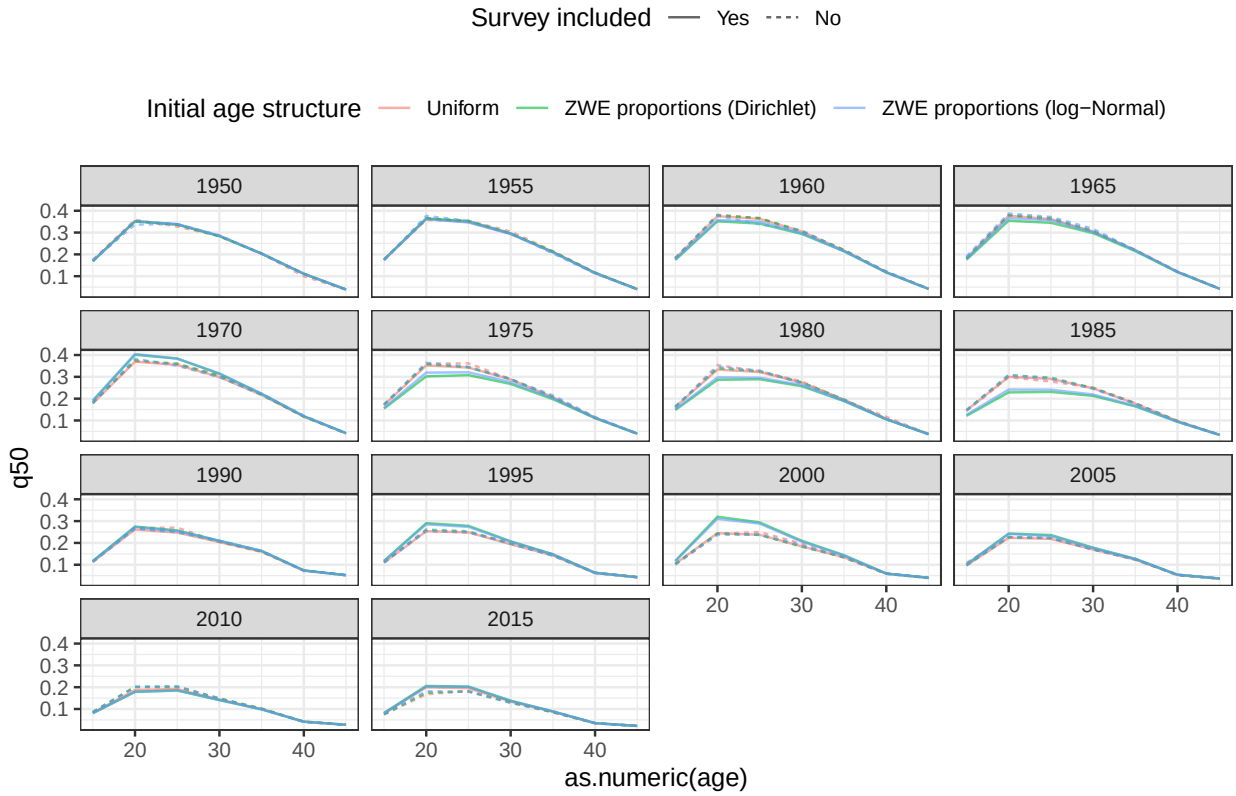


Figure 6: Estimated fertility rates under different initial population assumptions

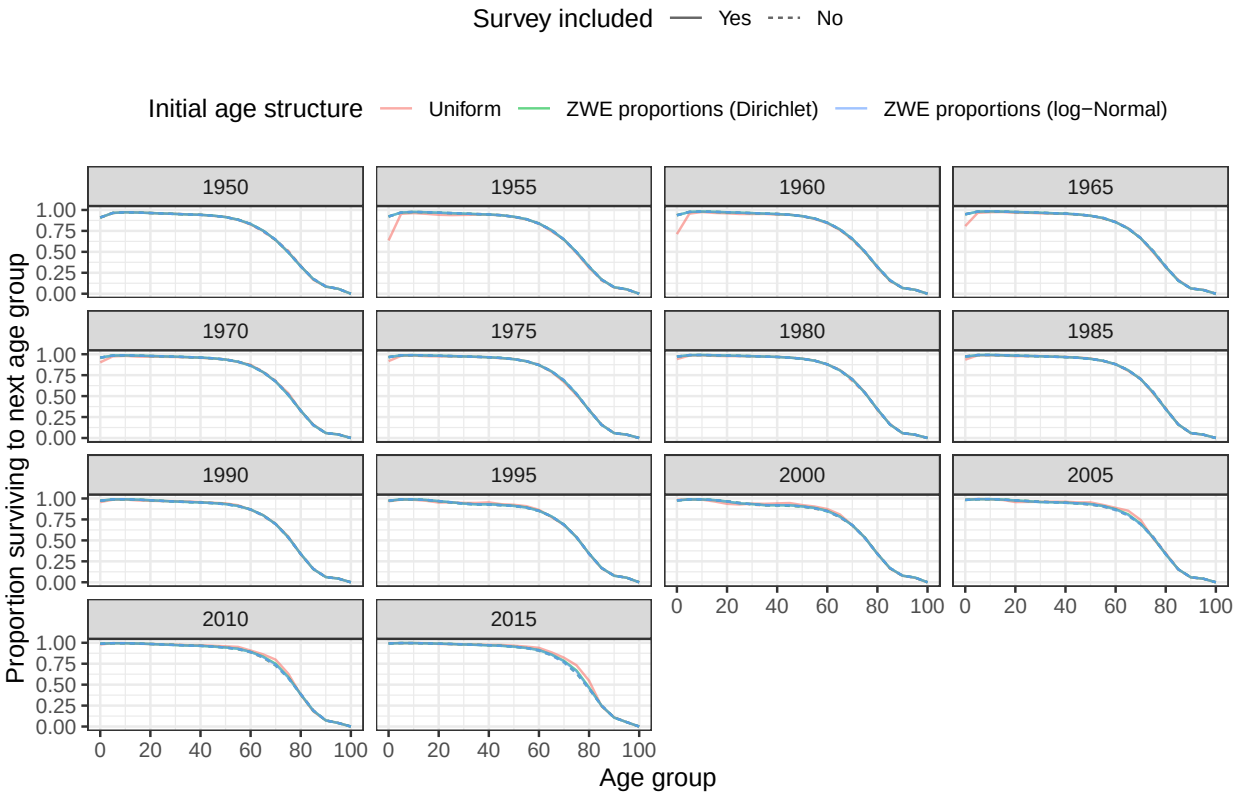


Figure 7: Estimated survival rates under different initial population comparisons

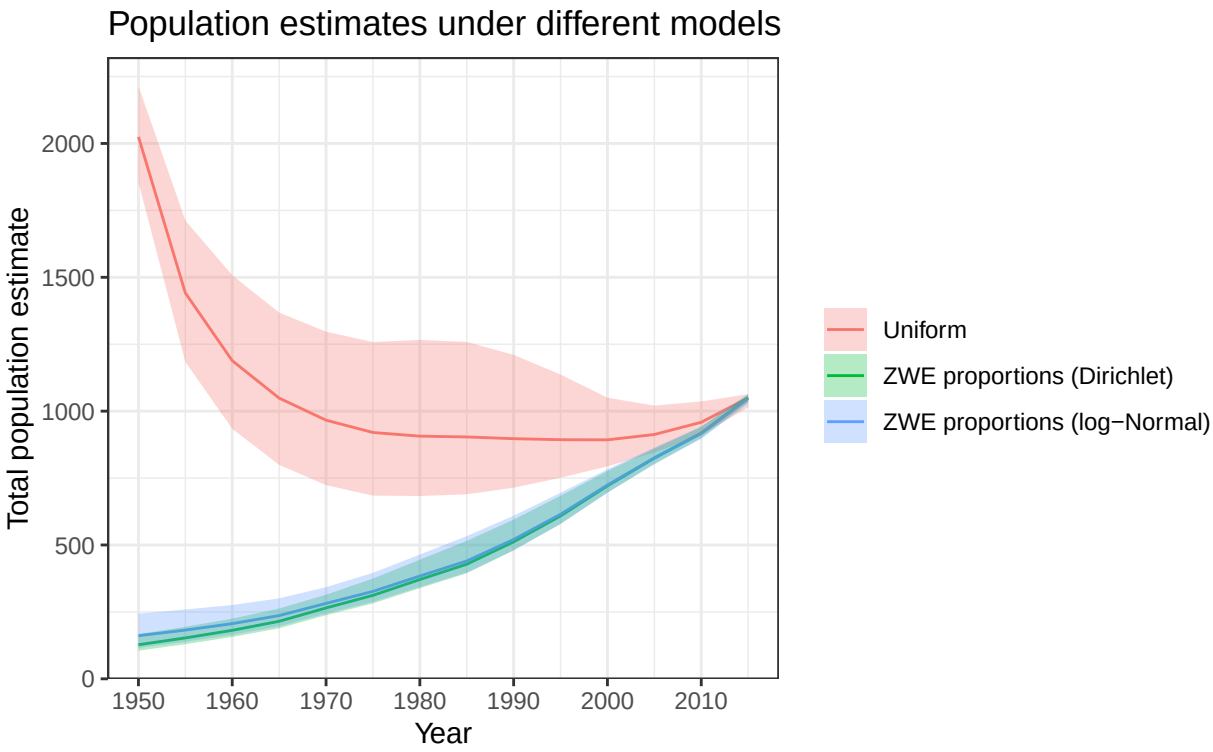


Figure 8: Estimated population sizes under different initial population assumptions.

Intuitively, this may also mean that incorporating population observations as independent, age-specific counts is more informative for the model than incorporating the observations separately as total counts and proportions. In practice, the latter option may be more applicable when, for instance, the total population count is considered reliable, but age information is not, since the uncertainty for the population total and age proportions are controlled separately.

2.3.2 Data incorporation

Figure 9 compares three variations on the data model, each using the Zimbabwean Normal log-ratio initialization. As expected, when the distribution is placed on the log totals, the estimate is relatively less constrained than in the other two alternatives.

We can see the differences in more detail in Figure 10. In particular, the right panel shows slight differences in the age structure. When the observation is incorporated as counts, the estimate follows the observed data very closely, whereas in the total/proportions model, the estimate is relatively smoother over the age structure and has wider uncertainty intervals.

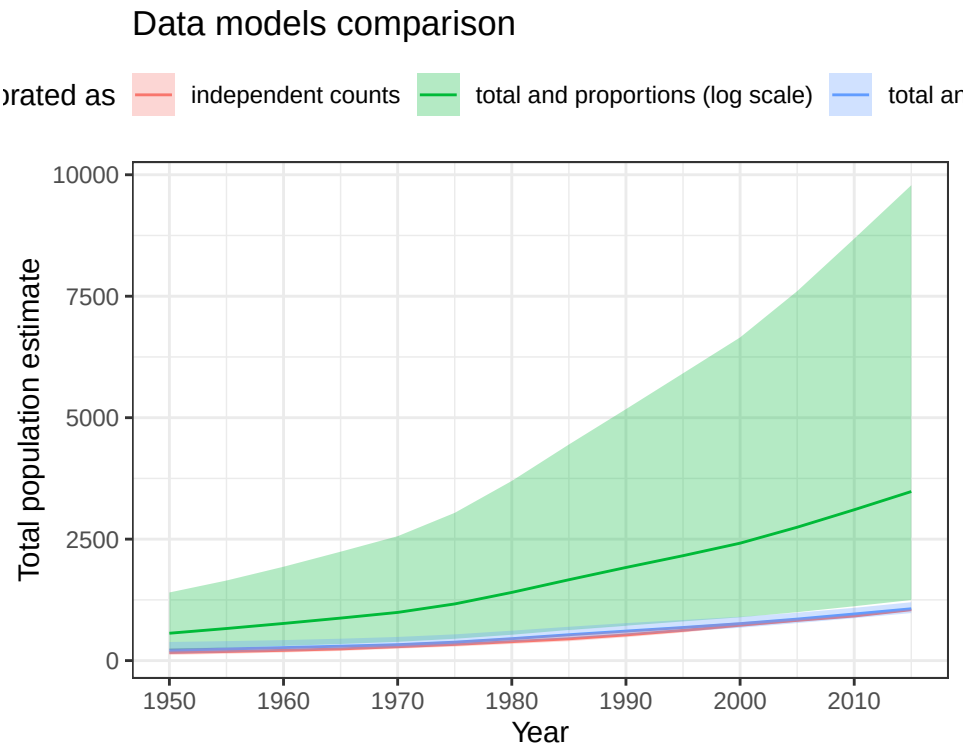
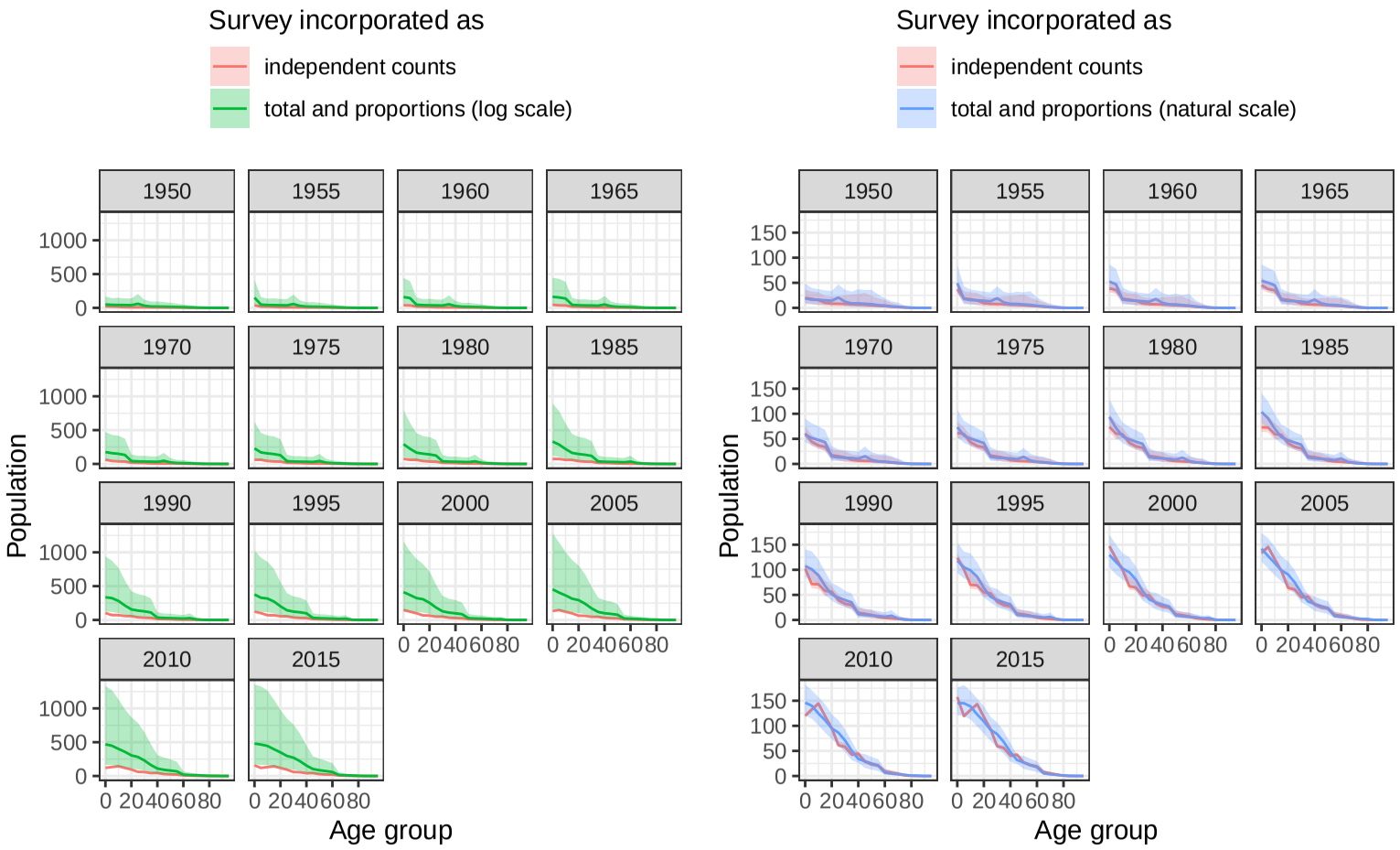


Figure 9: Total population estimates for the Shona population under different data models.



(a) distribution placed on log totals

(b) distribution of totals on natural scale

Figure 10: Comparison of population reconstruction using independent counts data model versus population total and proportions data models

2.4 Summary

In this section, we illustrated a population projection approach to estimating stateless populations using the Kenya Shona survey as an observation. We explored various ways to initialize a population using partial information, and several ways to incorporate an observation. Given the recency of the survey, we do not project forward from the time of the survey. However, this case study shows the implications of plausible model choices and when they may be suitable. Among models for age proportions, we found that the Normal log-ratio parameterization was more flexible and computationally stable. Using Hamiltonian Monte Carlo sampling, the Dirichlet distribution in this context is difficult to resolve, even with the Gamma parameterization. This is due to very small proportions at the older age groups.

Broadly speaking, model choices should be made according to data availability and reliability. For example, models that decouple population totals and age structure are particularly useful if age is unavailable or unreliable. If a total count from an observation is unreliable, then modeling it on the log scale may be appropriate to avoid overconstraining the model.

In this case study we applied a probabilistic demographic projection framework to back-project reliable survey data on the Shona population in Kenya. However, this framework is potentially useful in a number of other data availability settings, and we have developed the model set-up and code to be easily extended to other situations. For example, it would be possible to use this framework to forward-project historical counts, to include multiple sources of fragmentary data through time (either total counts or age-specific counts), and incorporate migration flows or other out-flows due to changes in legality. In general, the strength of the probabilistic projection framework is that the resulting estimates can be informed by multiple types of information, and incorporate different levels of uncertainty.

3 'Overlap' model proposal

In this section we discuss how to potentially leverage census data to produce estimates of stateless populations where other high quality data is not available. The central idea is behind this approach is this: although census data on stateless populations is generally acknowledged to be unreliable on its own, we can ground assumptions about the underrepresentation of stateless representation in censuses by looking at similar countries that have comparable census data and high quality data available. By calculating adjustment factors and assuming these hold to other, similar countries, unreliable census estimates could then be corrected for the underpresentation to produce a more reliable estimate.

We describe this approach below, but do not yet apply the estimation process due to lack of suitable data. The current data availability and what data are further needed for this method are discussed in the third section.

3.1 Method

The proposed method comprises three main steps. First, an adjustment factor for a (possibly outdated) census is estimated. The adjustment factor relates counts in a census (or stateless persons, or persons of unknown citizenship, for example) to a high quality data source on stateless counts. Second, the adjustment factor is applied to the surveyed population at the time of the census to estimate the stateless population. Lastly, the census-year-estimate of the stateless population is projected forward to present day produce an estimate of the current stateless population.

This model requires several data inputs. For some country of interest $c = c^*$, we must have some census, say, at year t^*, from which we can extract age-specific counts of respondents self-reported stateless (or of unknown citizenship). Assuming A age groups, we denote these counts by $\vec{y}_{c^*,t^*} = (y_{c^*,t^*,1}, \ldots, y_{c^*,t^*,A})$. Note that this census is for a country of interest for which we do not have another reliable stateless population counts.

We also require data from similar countries $c = 1, \ldots, C$ census data (comparable to that of c^*) is available in a year where reliable stateless data is also available. For country c, denote the most recent year where this overlap occurs as t^c, and denote the age-specific census counts $\vec{y}_{c,t^c} = (y_{c,t^c,1}, \ldots, y_{c,t^c,A})$ and denote the reliable stateless data as $\vec{z}_{c,t^c} = (z_{c,t^c,1}, \ldots, z_{c,t^c,A})$. This model may be able to be adapted to situations where some data is unavailable (e.g. the reliable data is not age-stratified), and we discuss possible modifications below.

Age-specific fertility and survival rates are required for population projection and denoted $f_{c,t,a}$ and $s_{c,t,a}$ respectively.

Finally, we require data or estimates of flows in and out of the stateless population of country c^*. Namely, if individuals are granted legal status, migrate elsewhere, or additional stateless individuals enter the population, then these should be incorporated in the projection step. If these quantities are negligible, or the census was taken very close to present day, then assuming them to be zero in the absence of data may not have a large impact on the resulting estimate.

3.1.1 Estimating the adjustment factor

Let ϕ denote the ratio of the number of individuals represented in the census data (y) to the number of individuals represented in the reliable data (z). For reference countries $c = 1, \ldots, C$, assuming that the age-specific data are available, we can calculate the ratios for each country-observation-age group, $\phi_{c,t,a} = y_{c,t,a}/z_{c,t,a}$.

In the simplest case, census data for all countries and the available high quality data all occur in the same year t, and we can model the distribution of ϕ as

$$\log \phi_{c,t,a} \sim N(\log \bar{\phi}_a, \sigma_\phi^2),$$

where $\bar{\phi}_a$ represents the mean ratio for age group a, and σ_ϕ captures the inter-country variability in ϕ. Note that this assumes that there are a set of countries similar enough in terms of likely census undercount that the mean adjustment factor is a reasonable estimate for countries without good quality stateless data.

The age-specific ratios $\bar{\phi}_a$ are assumed to have a smooth structure over age, which is enforced by expressing $\bar{\phi}$ using B-splines

$$\phi_a = \sum_{k=1}^{K} \beta_k B_k(a),$$

where $\{B_k\}_{k=1}^K$ represents a B-spline basis over the interval $[1, A]$, and β_k are coefficients to be estimated, assigned weakly informative priors

$$\beta_k \sim N(0, 100^2).$$

In principle the above setup can still hold as long as there is no systematic change in the censuses over time. In other words, if we believe that that the countries' censuses in question are reasonably comparable despite having occurred in different years in different countries,

then we do not require that all censuses happen at some time t. However, at this point we do still require that the census and reliable data for a given country are taken at the same time.

We may instead want to allow for the possibility that there is systematic change in the average quality of census data over time. In that case, the observations could inform a year-specific $\bar{\phi}_{a,t}$, such that for each observed $\phi_{c,t,a}$,

$$\log \phi_{c,t,a} \sim N(\log \bar{\phi}_{t,a}, \sigma_\phi^2),$$

with similar smoothing over age,

$$\phi_a = \sum_{k=1}^{K} \beta_{k,t} B_k(a),$$

and a random walk imposed on the spline coefficients to control the temporal variation

$$\beta_{k,t} \sim N(\beta_{k,t-1}, \sigma_\beta^2).$$

Modeling ratios over age would require much more data to obtain reasonable estimates.

3.1.2 Applying the adjustment factor

Once estimated from the previous section, the adjustment factor can then be applied to correct the census estimate of country c^*. Assuming that the census in countries $c = 1, \ldots, C$ is representative of the situation in c^*, we can estimate the true number of stateless $\hat{z}$, individuals in age group a at census time t^* as

$$\hat{\hat{z}}_{c^*,t^*,a} = (\hat{\phi}_{c^*,t^*,a})^{-1} \cdot y_{c^*,t^*,a},$$

where $\hat{\phi}_{c^*,t^*,a}$ comes from the predictive distribution,

$$\log \hat{\phi}_{c^*,t^*,a} \overset{\text{RNG}}{\sim} N(\log \phi_{t^*,a}, \sigma_\phi^2).$$

3.1.3 Projecting to current time period

After obtaining the adjusted estimate at the time of the census $\hat{z}_{c^*,t^*,a}$, we can project the population forward in time using the Leslie matrix approach described in the demographic projection section, and we give only a brief summary here. In short, the relevant fertility and

survival rates for some time t are structured in a Leslie matrix $L_{c,t}$ such that, when multiplied by an age-specific population vector $\vec{z}_{c,t}$, the resulting vector reflects the population after a fixed interval of time, accounting for the births and deaths in that period. The population $z_{c,t+1}$ at the next time step is calculated

$$\vec{z}_{c,t+1} = L_t \cdot \vec{z}_{c,t}.$$

In cases where citizenship is granted *jus soli*, the "fertility" terms in L_t can be set to zero to reflect the fact that individuals born in that country are not stateless.

Depending on the context, we may also incorporate inflows and outflows of this population, for instance if individuals are known to have been granted legal status. We denote the net inflow as $\vec{m}_{c,t}$. To approximate flows happening throughout the interval, we assume that half of the flow occurs at the beginning of the interval (thus experiencing births and deaths) and half at the end, such that the population after accounting for migration is

$$\vec{z}_{c,t+1} = L_t \cdot \left(\vec{z}_{c,t} + \frac{\vec{m}_{c,t}}{2}\right) + \frac{\vec{m}_{c,t}}{2}.$$

For simplicity of notation, we denote this projection operation $p_t(\cdot)$ such that

$$z_{t+1} = p_t(z_t).$$

Iteratively applying this process allows us to project the population forward in time as necessary to obtain a present-day estimate.

One challenge of this approach, is the assumption of demographic rates and the uncertainty surrounding them. We may, for instance, set the rates at the WPP estimates which are readily available. However, prior predictive checks may be needed to calibrate the uncertainty to reflect a reasonable range of rates.

3.1.4 Adaptations for partial data

There are possible adaptations that can be made if full, detailed data are unavailable. The approaches described below require additional assumptions, and so in general it is always better to have more detailed data.

3.1.4.1 Age breakdown unavailable If the reliable data for some country c_1 has a count $N_{c_1,t}$, but are not disaggregated by age, possible options are to use the age structure

from other high quality data. This assumes that the age-structured data approximates the true age structure of the stateless population, which may make this particularly applicable when the stateless population largely comes from one place of origin, and there are detailed data about the origin population.

The approach is similar to that of using the Zimbabwean population proportions to initialize the Kenya Shona population projection. Let $(p_{c,1}, \ldots, p_{c,A})$ denote the age proportions of the population in country c. Choosing some age group a' to act as a reference category, we calculate log-ratios of model proportions and denote these using ρ:

$$\rho_c = (\rho_{c,1}, \ldots, \rho_{c,A}) = \left(\log \frac{p_{c,1}}{p_{c,a'}}, \ldots, \log \frac{p_{c,A}}{p_{c,a'}} \right).$$

We can then model the log ratios as

$$\rho_{c,a} \sim N(\bar{\rho}_a, \sigma_\rho^2)$$

and the age structure for country c_1 can be estimated

$$\hat{\rho}_{c_1,t,a} \overset{RNG}{\sim} N(\bar{\rho}_{t,a}, \sigma_\rho^2)(\hat{p}_{c_1,t,1}, \ldots, \hat{p}_{c_1,t,A}) = \frac{(\exp(\hat{\rho}_{c_1,t,1}), \ldots, \exp(\hat{\rho}_{c_1,t,A}))}{\sum_{i=1}^{A} \exp(\hat{\rho}_{c_1,t,i})}$$

Age specific counts can then be estimated

$$\hat{z}_{c_1,t,a} = \hat{p}_{c_1,t,a} \cdot N_{c_1,t},$$

and can then be used in the estimation process described above.

3.1.4.2 Census and reliable data exist but are not in the same year In the case that some country c has census data $\vec{y}_{t_1}$ available at year t_1 and reliable data $\vec{z}_{t_2}$ available at year t_2, we may apply a projection step before we calculate the adjustment factor. In either case, the adjustment is performed on the (forward or backward) projected population.

If the reliable data was recorded first ($t_2 < t_1$) then we first project forward in time. Recalling that p_t denotes the projection operation accounting for population flows, the stateless population at census time t_1 is

$$\vec{z}_{t_1} = p_{t_1-1}(p_{t_2-2}(\cdots p_{t_2}(\vec{z}_{t_2}))).$$

We can then calculate the adjustment factors $\phi_{t_1,a}$ at the time of the census

$$\phi_{t_1,a} = \frac{y_{t_1,a}}{z_{t_1,a}}.$$

On the other hand, if the census data was recorded first ($t_1 < t_2$), we model ϕ similarly at the time of the census. However, the projection is done starting at time t_1 using the adjusted census count $\vec{y}_{t_1}/\vec{\phi}_{t_1}$, and we iteratively apply the population projection until time t_2

$$p_{t_2-1}\left(p_{t_2-2}\left(\cdots p_{t_1}\left(\frac{\vec{y}_{t_1}}{\vec{\phi}_{t_1}}\right)\right)\right) = \vec{z}_{t_2}.$$

3.2 Data availability

As mentioned above in Section 3.1, this method requires:

1. for the country of interest (where no good quality data exist), census data which is to be adjusted,
2. for a selection of "similar" countries, comparable census data and a reliable observation,
3. age-specific fertility and survival rates for each country,
4. "external" flows in and out of the stateless population (not due to birth or death) for each country, from the time of the earliest observation used for that country

Thus far, for various reasons, we do not feel confident applying this approach to any countries. To our knowledge, there are few countries that have data considered reliable or plausibly reliable, and even fewer among them which have census data readily available on IPUMS. We discuss select cases below.

In Europe, Ireland is the only country reporting census data on IPUMS and with plausibly reliable data on stateless persons in the same year. The 2011 census in Ireland allows for respondents to self-declare unknown citizenship and no nationality, and both responses are available on IPUMS. For the 2016 census, only the unknown citizenship response is available on IPUMS. UNHCR data for Ireland are consistently available starting from 2015. Other European country-years where the adjustment may be applied (i.e. those with census data available with IPUMS with unknown/stateless responses, and no reliable estimate) include Austria 2011, Belarus 2009, Greece 2011, Italy 2001, Poland 2011, and Portugal

2011, Romania 2011, and Slovenia 2002. However, since the available census are at least 10 years old, we would not be able to account for external flows over this period.

A high quality estimate of stateless persons exist in Thailand, and the 2000 census reports respondents with unknown citizenship. However, previous data is likely not of similar quality, and we do not have data available regarding external flows. The situation is similar in Malaysia which also had a 2000 census. Results from these two countries could inform others in the region with available census data, such as Indonesia 2010 and Laos 2005.

Kyrgyzstan's 2009 census allowed for stateless responses, and the most recent observation is considered reliable. If UNHCR figures from this time are also reliable, then the method could be applied here. However, population projection from the more recent UNHCR figure is not an option here without data on how many persons gained legal status in the time since the census.

It is worth noting that some censuses are not available on IPUMS and were therefore not part of our search. In particular, if more recent censuses are available, then the problem of missing data on flows is not as severe.

For a full graphical summary of data availability and overlap, please see the `data_availability_explore` document in the GitHub repo.

3.3 Summary

In this section we discussion a conceptual framework for estimating stateless populations based on census information, as well as adjustment factors calculated based on other, similar, countries. In an ideal situation for this model approach, we would have many recent censuses available with stateless information (or other data on unknown citizenship, for example) for countries where we also have reasonable stateless counts from other data sources. This would allow us to calculate a number of different adjustment factors that could then be applied to other countries. However, based on our data exploration (relying solely on censuses on IPUMS), there are very limited situations where there is an overlap of information, particularly in recent time periods. While, it is possible that censuses may be available from other sources that may be updated more quickly than IPUMS, in general for this method to be used, it may need to be combined with demographic projection techniques (as discussed for the Shona population) in order to project historical adjustment factors forward.

4 At-risk populations case study: estimation in the United States

This section discusses the case study of estimating stateless populations in the United States. We use data from the American Community Survey to extract information about the size of immigrant populations that are potentially at risk of statelessness. We then describe trends and patterns in these data and illustrate how population size can be estimated and projected over time with uncertainty.

As discussed in more detail below, one of the main drawbacks of this approach is that estimates are of those populations potentially at risk of statelessness, rather than to be known to be stateless. We thus also illustrate a modeling strategy that incorporates the probability of an individual with certain characteristics is stateless. This approach is illustrated in adjusting the Nepal immigrant population by their year of arrival.

4.1 Data

We used data from the 1-year American Community Surveys (ACS) over the period 2005–2019. The ACS is a nationally representative survey that covers roughly 1 percent of the US population. Each year, it gathers detailed data for all states and for cities, counties, metropolitan statistical areas, and Public Use Microdata Areas (PUMAs), a US Census–defined geographic area of 100,000 people or more.

Data were extracted from the US Census Bureau's API using the `R` statistical language and the `censusapi` package. We only considered sample respondents who foreign-born and were not US Citizens. We extracted variables on age, sex, place of birth, nativity, ancestry, year of entry, household language, nativity of parents, and state of residence.

4.1.1 Populations at risk of statelessness

To identify these groups of populations that are potentially stateless or potentially at risk of statelessness in the United States, we largely followed the approach of Kerwin et al. (2020). In their paper, Kerwin et al. developed at-risk profiles based on reviewing UNHCR 'Mapping Statelessness' reports, and other consultations and interviews.

The following is a list of all profiles considered. For more information on how these groups can specifically be identified using the variables available, see the `at_risk` R script. Note

that based on the information available not all groups could be identified in the ACS, and these are flagged below.

4.1.1.1 Europe and Eurasia

- Emigrants from the Former Soviet Union before Its Collapse Who Arrived in the US before 1992
- Ethnic Russians, Belarussians, and Poles from Latvia
- Members of Ethnic Minority Groups from Lithuania
- Armenians from Azerbaijan
- Azerbaijanis from Georgia
- Meskhetian Turks
- Roma, Ashkali, and Balkan Egyptians
- Individuals with Yugoslavian Passports Who Entered the US before 1992
- Born in North Macedonia, Other Ex-Yugoslav Descent
- Born in Croatia, Serbian Descent
- Roma Born in Italy and Germany

4.1.1.2 Middle East and North Africa

- Syrian Refugee Children Born Abroad
- Feyli Kurds from Iraq. Cannot be identified from non-Feyli Kurds
- Syrian Kurds
- Lebanese Kurds and Bedouin. Cannot be identified
- Tebu Libyans. Cannot be identified
- Palestinians
- Bidoon

4.1.1.3 Asia and South Pacific

- Nepalese Born after 1990
- Ethnic Nepalis (Lhotshampas) Born in Bhutan
- Rohingya. Cannot be identified
- Other Minorities from Myanmar
- Hmong from Laos
- Hmong from Thailand

- Members of Thai Hill Tribes
- Thai-Born Children of Burmese Refugees
- Chinese without Hukou Registration
- Tibetans
- Afghan Jogi. Cannot be identified
- Stateless Groups from India. Cannot be identified
- Bengalis from Pakistan
- Ethnic Minorities from Malaysia
- Ethnic Vietnamese and Khmer Krom from Cambodia
- Chinese Cambodians from Vietnam. Cannot be identified
- Stateless Persons from Brunei. Cannot be identified

4.1.1.4 Sub-Saharan Africa

- Ivoirians with Ancestry in Mali, Guinea, and Burkina Faso. Cannot be identified
- Stateless Groups from Kenya. Shona can only be identified after 2016
- Karana of Madagascar. Cannot be identified
- Bakassi of Nigeria and Cameroon. Cannot be identified
- Zimbabweans with Origins in Neighboring Countries. Cannot be identified
- Returned Mozambican Refugees in Mozambique. Cannot be identified
- Black Mauritanians. Cannot be identified
- Sahrawi (Born in the Western Sahara or Algeria). Cannot be identified
- Ethiopians with Eritrean Ancestry
- Eritreans with Ethiopian Ancestry
- Sudanese-Born Individuals of Dinka and Nuer (South Sudanese) Descent, Arrived in 2011 or Later. Cannot be identified
- Born in South Sudan, Main South Sudanese Ethnic Groups, Arrived in 2011 or Later. Cannot be identified
- Banyarwanda and Banyamulenge from the Democratic Republic of Congo

4.1.1.5 Americas

- Dominicans (Dominican Republic) of Haitian Ancestry
- Bahamians of Haitian Ancestry

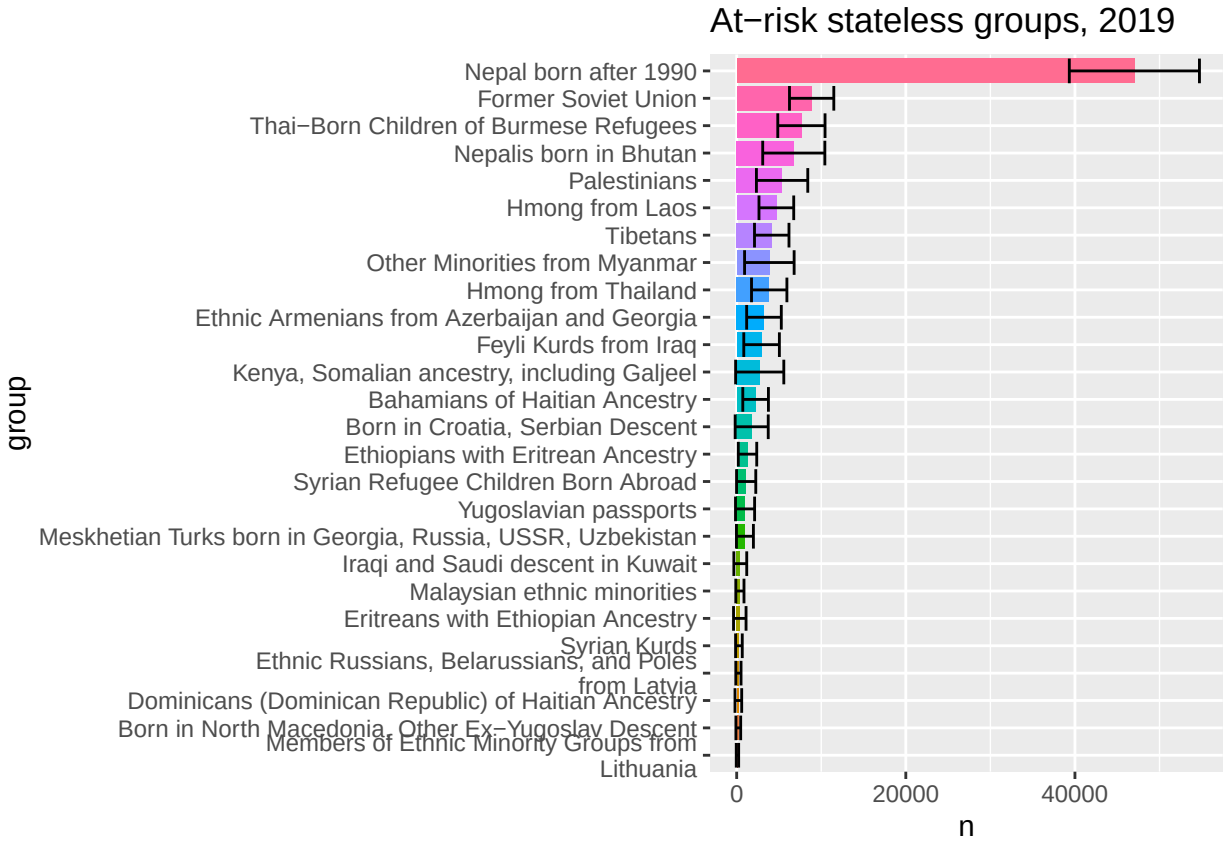
4.2 Description of data

We now illustrate main patterns in the data extracted by profile group, age, sex, and region. Estimates are calculated based on person-weights and standard errors around the estimates are calculated based on replicate weights provided in the ACS data. Note that estimates of particular groups are often based on very small samples in the ACS and so standard errors are large.

4.2.1 Totals by group

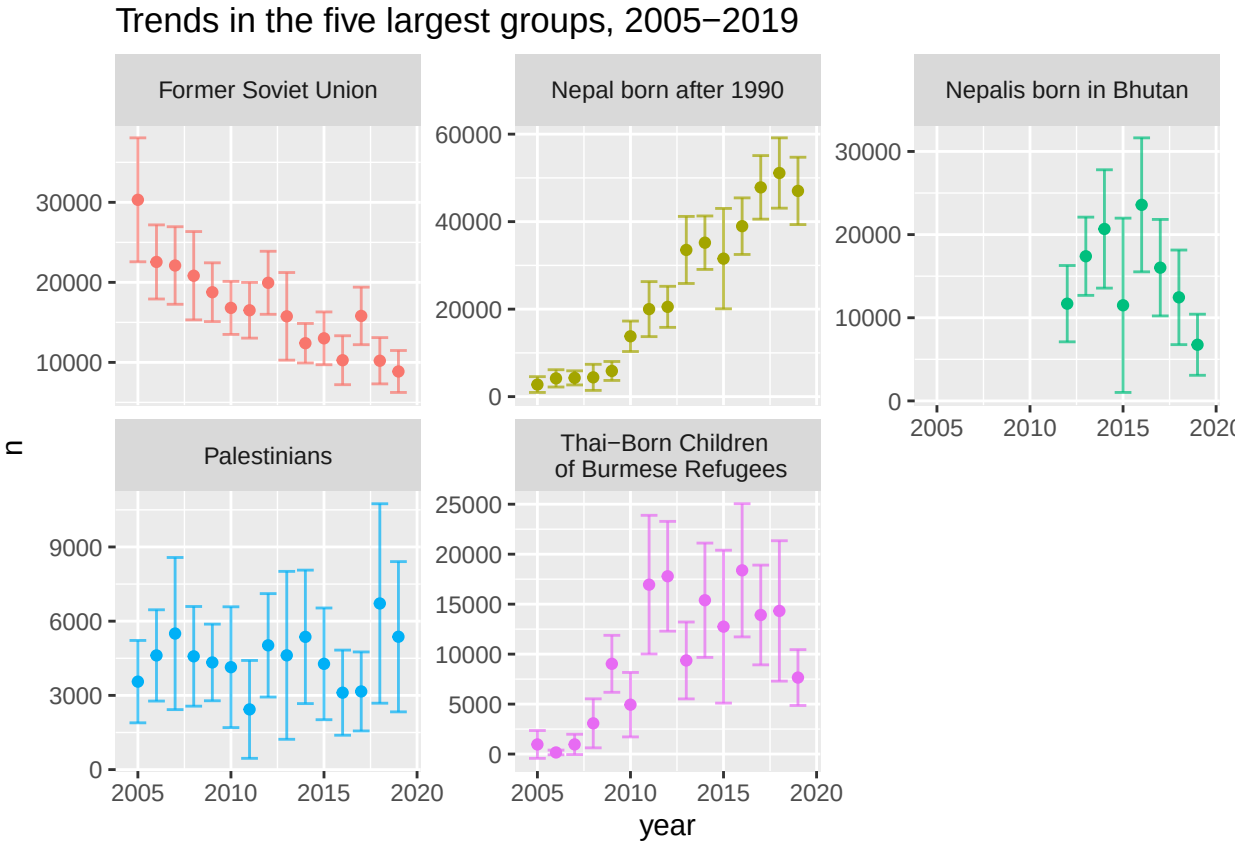
Based on the 2019 ACS, there were an estimated 111621 (CI: 99327, 123915) persons identified in the at-risk profiles listed above.

The graph below shows the estimated number of persons in each statelessness risk profile in 2019. By far, the largest identified group are those born in Nepal after 1990, who constitute almost 50% of the total number of persons identified.



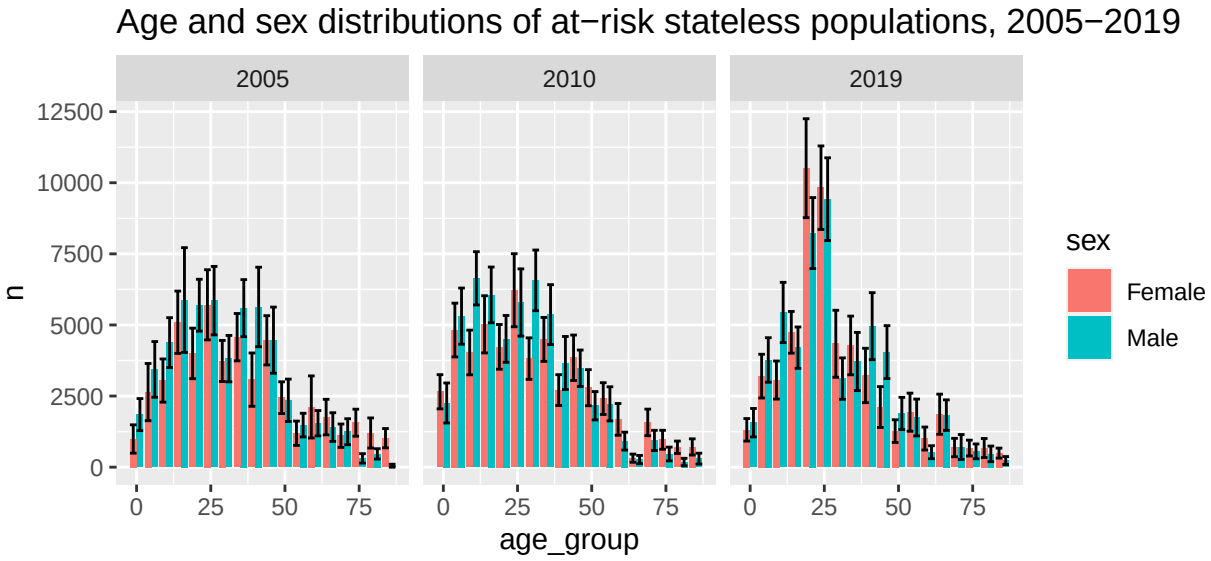
Looking at the trends over time in the five largest groups, the population of those from the Former Soviet Union has unsurprisingly decreased over time. In contrast, the population of those from Nepal born after 1990 has steadily increased. Note the uncertainty in the

estimates from the raw ACS data is quite large, even for these five largest groups. For example, the uncertainty in the 2019 number of Palestinians spans almost 3,000 people.

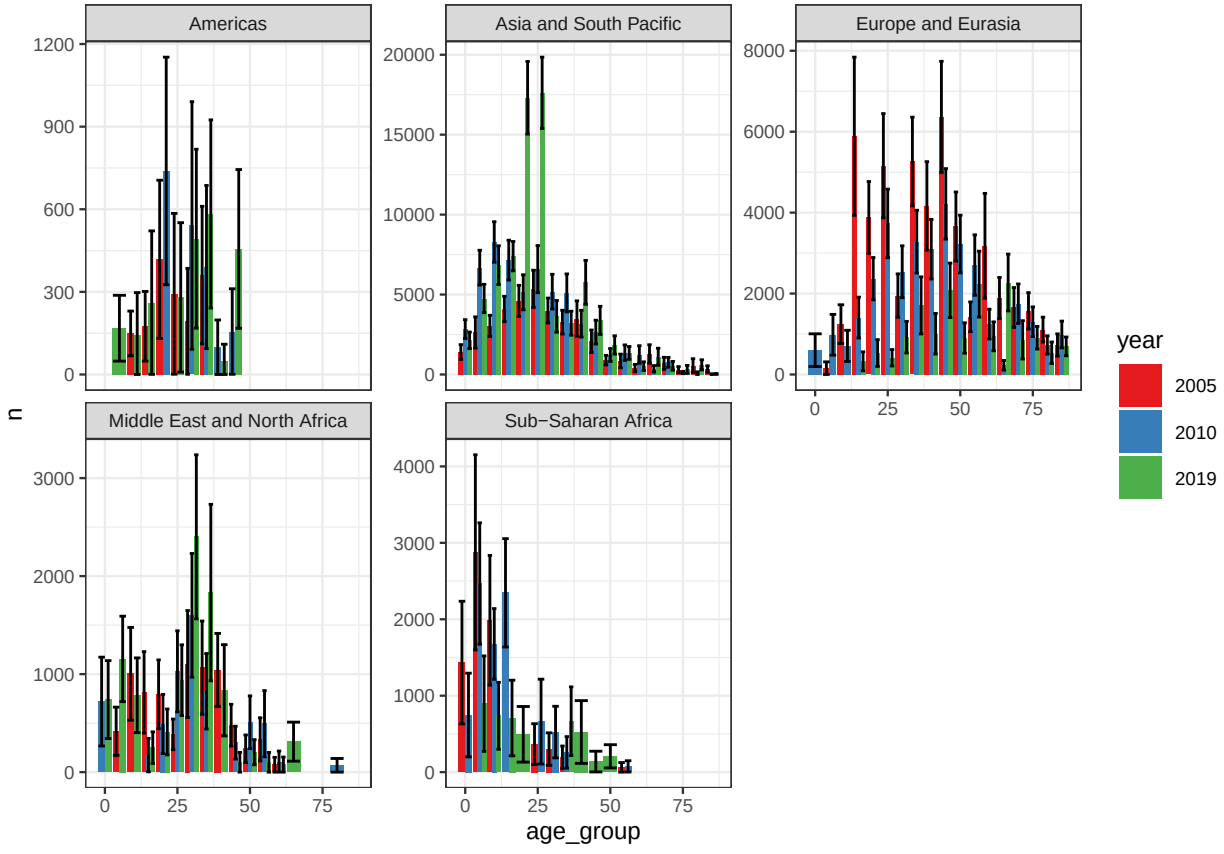


4.2.2 Totals by age/sex

Looking at the trends in the age/sex distributions of the total group at risk to statelessness, the figure below shows that the age distribution peaks at ages 20-25, and this peak becomes more pronounced over time. In 2019, 19249 (CI: 14600, 23898) persons aged 25-29 were at risk of statelessness. In general, there are more men than women at risk of statelessness.



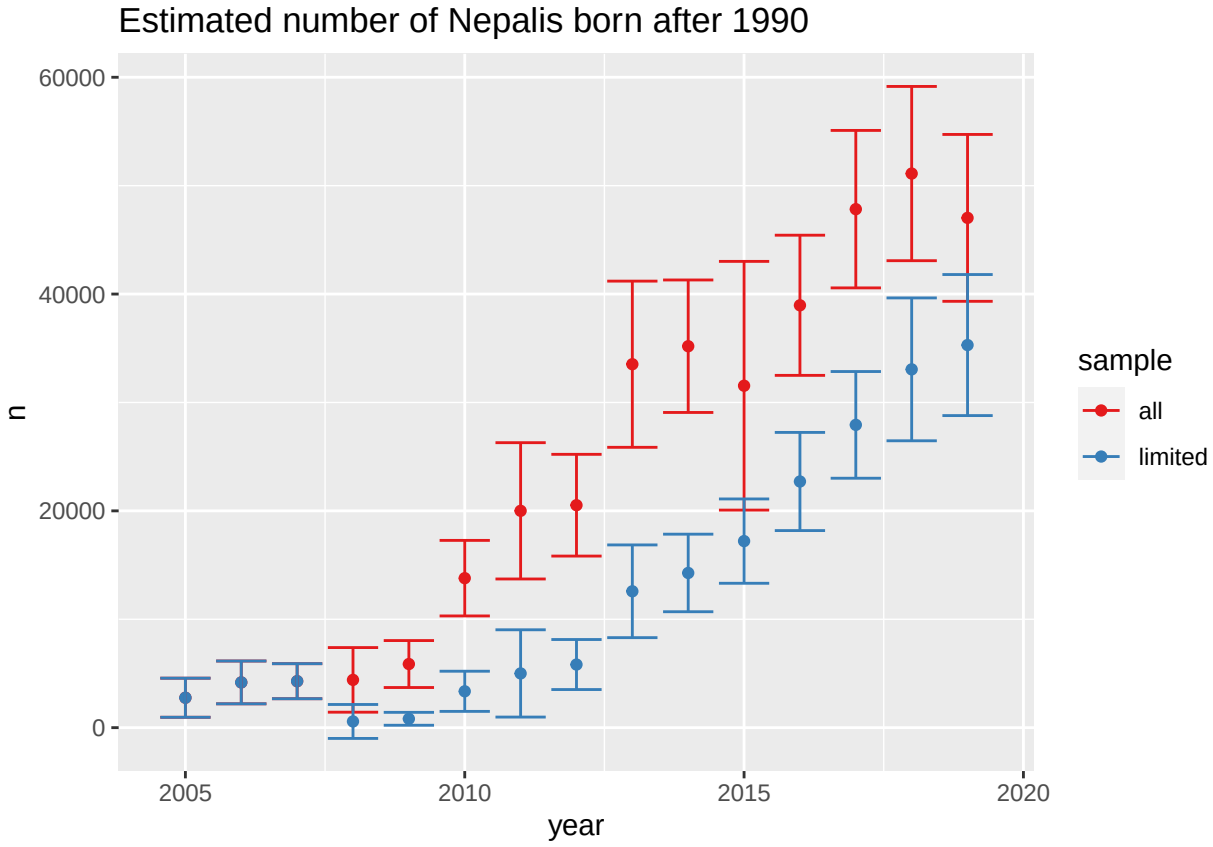
Looking at age distributions by region, we can see that the large peaks in 20-30 age groups in 2019 are a consequence of those from the Asia and South Pacific region, and in particular from Nepal. The age distribution in Europe and Eurasia, which is driven by immigrants from the Former Soviet Union, is relatively older and aging over time.



4.2.3 Restricting Nepalese based on nativity of parents

The estimate of the total population of those potentially at risk to statelessness from the ACS is dominated by those born in Nepal after 1990, which was estimated to be around 47028 people in 2019. However, it seems unlikely that this entire population is indeed stateless. In a 1990 revision of its constitution, Nepal restricted automatic citizenship to those descended from a Nepalese father. Citizenship rules changed again in 2006, and citizenship on the basis of birth became possible if individuals applied within 2 years. Finally, in 2011, Nepal passed further reforms that allow children to acquire citizenship through mothers if their father is unknown or absent. In practice it may be difficult, particularly for single women, to register their children as citizens.

The ACS provides additional information which may help to further refine this profile to better reflect those who are actually stateless. In particular, for those respondents who still live with their parents, it seems likely that those at most at risk of stateless are respondents who do not live with a father, and their mother is not native born. As shown in the figure below, the restriction decreases the number of people in this group by around 8000 in 2019.



4.3 Time series model

As seen above, the observations of populations by profile group are quite noisy over time. To aid in estimation within the period of observation, as well as forward projection to future time points, we modeled each profile group with a Bayesian time series model.

4.3.1 Model

Let $y_{g,t}$ be the number of persons in profile group g in year t. We model these on the log scale as

$$\log y_{g,t} \sim N(\mu_{g,t}, s^2_{g,t})$$

where $\mu_{g,t}$ is the expected number of persons in profile group g in year t and $s^2_{g,t}$ is the corresponding sampling variance calculated based on the ACS. We then model the expected numbers $\mu_{g,t}$ as a second-order random walk:

$$\mu_{g,t} \sim N(2\mu_{g,t-1} - \mu_{g,t-2}, \sigma^2_g)$$

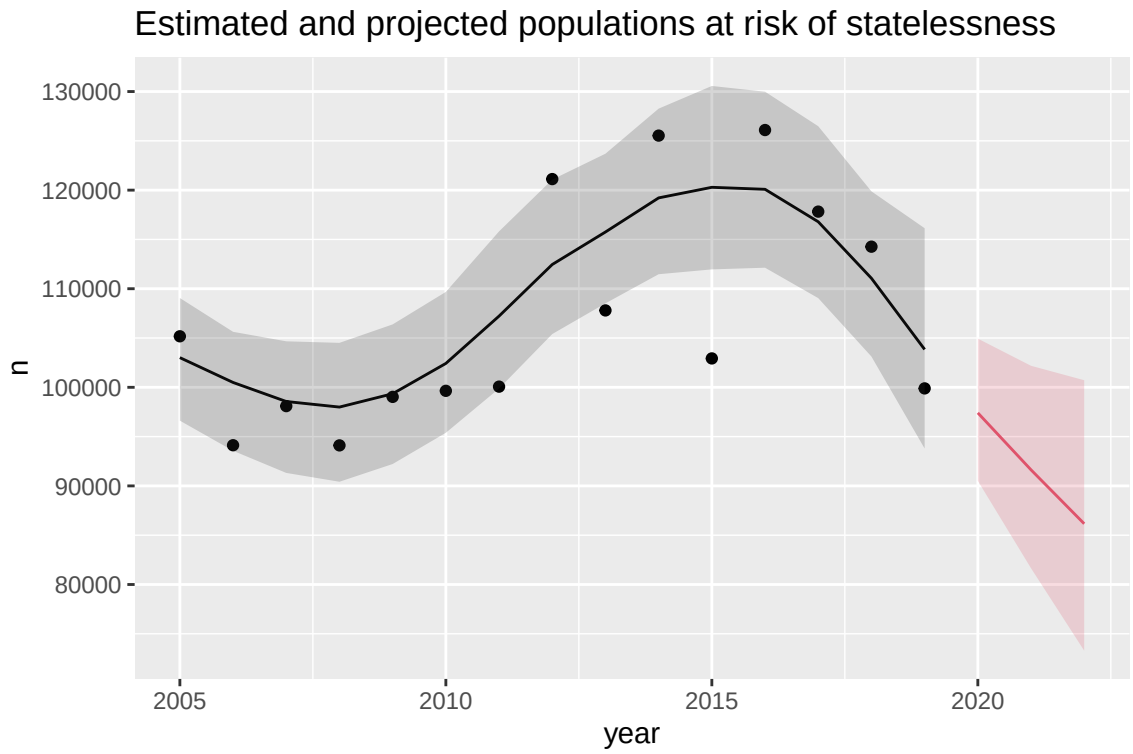
This model penalizes the second-order differences in the trajectory of the expected numbers over time. This is equivalent to penalizing fluctuations away from a linear trend. There is a separate variance term σ^2_g for each profile group, allowing for different variability over time for each group. However, these variance terms are modeled hierarchically such that

$$\log \sigma_g \sim N(\phi, \tau)$$

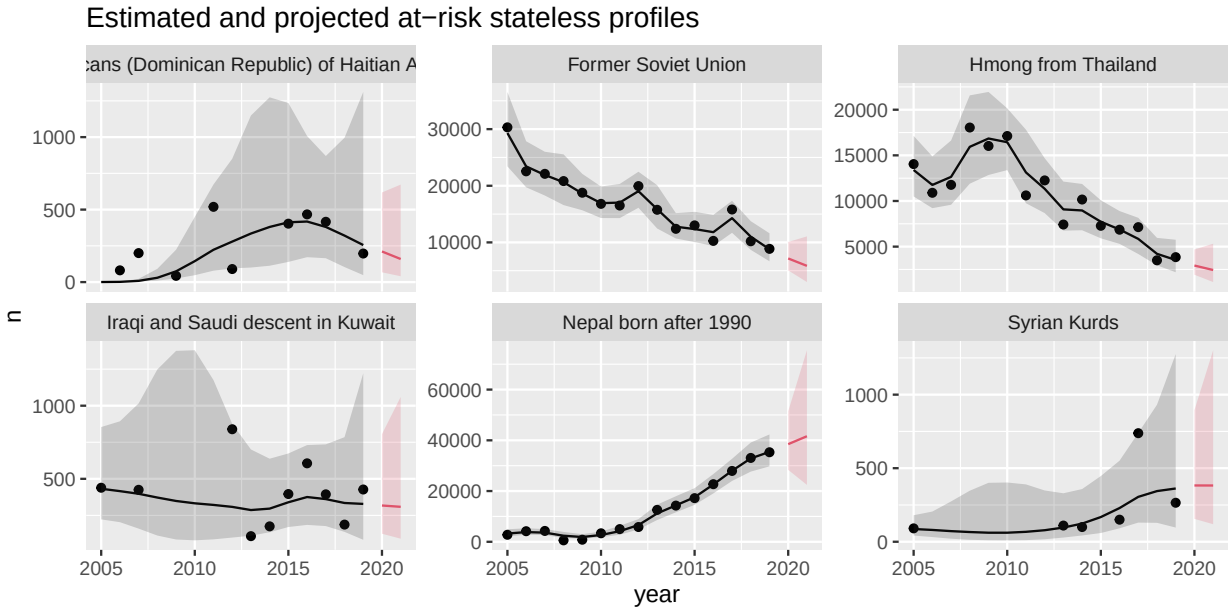
In this way, information about the variability in each time series is shared across groups, and so those groups with missing observations (and thus less information) are partially informed by other groups.

4.3.2 Results

The figure below shows the estimated and projected total population identified to be at risk of statelessness. If recent declines continue, it is estimated to be approximately 86000 persons at risk of statelessness in 2021.



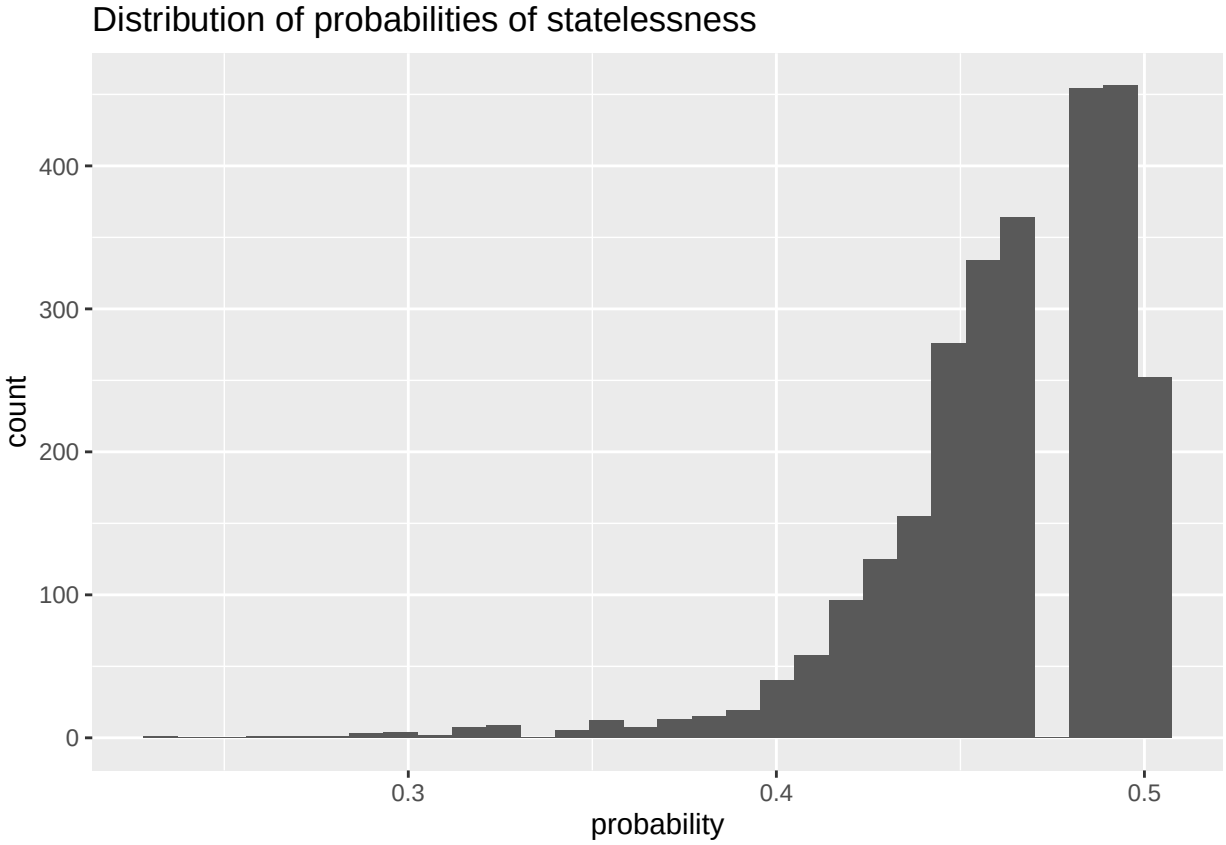
Looking at specific groups, the figure below illustrates estimates and projections for six different profiles of varying sizes and trends. Note that the smaller population groups, for example Syrian Kurds, have much larger uncertainty around estimates compared to the large population groups, such as those from the Former Soviet Union.



4.4 Adjusting probabilities of statelessness

It is impossible to know the number of stateless persons from just the information provided in the ACS alone; the best we can do is get an estimate of the population at risk of statelessness. However, if we had additional information on the likelihood of statelessness based on various individual characteristics then we could incorporate this information into the above model.

In this section we illustrate this idea with an application to the Nepal profile group. There are many potential characteristics that could be associated with the likelihood of being stateless; for example education, English language ability, marital status, etc. Here we assume that the number of years resided in the US is inversely related to the probability of being stateless. In the absence of any other information, we assume that there is a 50% chance of being stateless in the year of arrival, and that the probability decreases for ever year in the US, down to a minimum of 5%.[2] The graph below shows the distribution of these implied probabilities based on all ACS observations in the sample.



This varying probability of statelessness can be incorporated into the time series model stochastically, taking into account the uncertainty in the probability assignment. In

[2]This is a completely arbitrary function in the absence of any other information; ideally this could be informed by other data sources or expert knowledge.

particular, let z_i be the probability that individual i is stateless. This is modeled as a Bernoulli distribution,

$$z_i \sim \text{Bernoulli}(p_i)$$

where p_i is the probability of statelessness, and is equal to

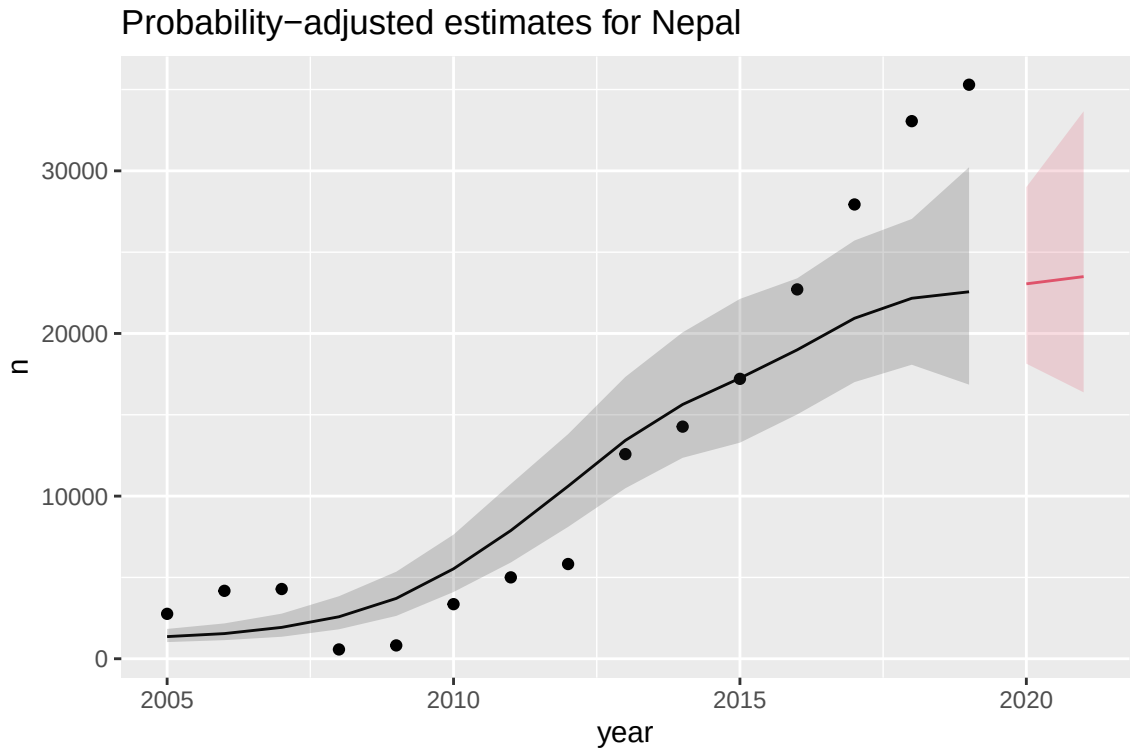
$$p_i = \max(0.05, 0.5 - 0.01x_i)$$

where x_i is the number of years that individual i has been in the US. The sum of migrants in a particular year t is then equal to

$$y_t = \sum_{i \in t} z_i$$

And y_t can then be modeled as a time series as shown in the previous section.

The results of this model are shown below. For comparison, the raw ACS data are plotted as the black dots. This illustrates that estimates and projections are substantially lower in the more recent years.



4.5 Summary and limitations

The American Community Survey provides a good quality data source on the number of people at risk of statelessness in the United States, based on pre-defined profiles. However, there are several limitations to using these data. Firstly, there is no way of knowing whether those at risk of statelessness are stateless, just using the ACS data alone. Ideally, we would have another data source that captures stateless persons more directly, in which we could compare to the ACS to calibrate the results. In lieu of this, as shown above, it is possible to model the probability of persons with different characteristics being stateless, accounting for the uncertainty in a fully stochastic framework. However, for this strategy to be the most effective, we would also need additional data sources and the likelihood of at-risk groups being stateless or becoming citizens.

The second major limitation is that the ACS does not capture all known profiles of those persons who are at risk of statelessness. For instance, there is no way of identifying Rohingya, or at-risk stateless groups that are specific ethnic minorities within some countries. In the Kerwin et al (2020) paper, they supplement the ACS data with data from WRAPS, which captured some of the groups that were known to be missing in the ACS. Unfortunately, the data they presented in that paper are no longer available at that level of granularity due to changes in the legislation during the Trump administration.

Notwithstanding, the ACS and other similar high-quality surveys in other countries (such as HILDA in Australia) show potential to help try and estimate those at risk of statelessness. However, for these estimates to be more accurately capturing the true number of stateless people, we would need additional data sources, most likely administrative, to calibrate and adjust the larger population.

5 Conclusion

This report outlined three different approaches to possibly estimate stateless populations in different contexts. The first section outlined a demographic projection framework, which was applied to data on the Shona population in Kenya, but can be extended and applied in other situations where it would be useful to project populations backwards and forwards in time, taking into consideration different data sources. The second section proposed a general approach to leverage census information across a number of different countries. The motivation behind this proposal was the assumption that the degree of underreporting of statelessness in some countries tells us something about the degree of underreporting in other, similar

countries. The last section outlined how ACS data can be used to estimate populations at risk to statelessness over time, and proposed a probabilistic method to incorporate the probability that an individual with certain characteristics is stateless at a given time point.

While the approaches presented all show potential in helping to estimate stateless populations, we were unable to produce reliable estimates for a wide range of countries, largely due to data availability issues. In particular, the 'overlap' method, which perhaps is the mostly widely applicable method (requiring relatively less data), was not run on actual data due to the limited number of countries that had recent censuses available on IPUMS. Going forward, we believe this work and the methods discussed here would benefit most from collaboration with substantive area experts who could assist with data sourcing and other knowledge that could help to inform model parameters.

References

Kerwin, D., Alulema, D., Nicholson, M., & Warren, R. (2020). Statelessness in the United States: A Study to Estimate and Profile the US Stateless Population. Journal on Migration and Human Security, 8(2), 150-213.

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(501), 96-110.