Small area estimation of child mortality in sub-Saharan Africa: Administrative level 2

Jessica Godwin[1] and Jon Wakefield[1,2]

[1]Department of Statistics, University of Washington
[2]Department of Biostatistics, University of Washington

1 Introduction

The under-five mortality rate (U5MR) is an important indicator of overall mortality levels and health in a population. The UN's former millennium development goals (MDGs) and current sustainable development goals (SDGs) both introduce targets of reduction in U5MR for countries. The ability to make estimates at a finer administrative level than the national level allows for identification of small areas that are relatively slow or fast in their reduction of child mortality. This can be useful for governments and organization trying to efficiently allocate resources and implement interventions.

Mercer et al. (2015) develop a method of space-time smoothing of small area estimates of child mortality that account for the complex survey design in DHS data. However, they implement this method at the Admin-1 level. We extend this method to make estimates at the Admin-2 level (a finer subdivision of a country) and to also include summary birth history data from MICS surveys and censuses. Most DHS surveys are stratified at Admin-1 level, which leads to low sample sizes for some Admin-2 areas. With low sample sizes, we often see no deaths in a survey. This leads to a point estimate for that area of zero, which we do not believe. We develop a method for using the neighboring Admin-2 areas to provide an adjusted estimate of child mortality in the Admin-2 area. Additionally, we incorporate summary birth history data (SBH) from MICS and censuses. We develop a model that allows for smoothing estimates at a yearly time scales from SBH data and estimates at a coarser time scales from FBH data. Here we apply our method to the country of Kenya, but will proceed with fitting this method to all countries in sub-Saharan Africa with geolocated DHS data using MICS and census data where applicable.

2 Methods

2.1 Data

We use full retrospective birth histories (FBH) from DHS surveys that cover the period of estimation 1980-2015 and have GPS location of clusters available. The geolocation data is necessary, as most DHS surveys are stratified at Admin-1 or coarser levels and urban/rural classification. The GPS locations of sampled clusters allow us to assign those clusters to the finer Admin-2 area. For Kenya, this includes the 2003, 2008 and 2014 DHS. (It is worth noting that there are 8 provinces and 47 counties in Kenya. Until recently, Admin-1 was the 8 provinces, now it is the 47 counties. The 2003 and 2008 surveys are stratified on the 8 provinces, so we treat the 47 counties as our new, finer level of granularity.) We use MICS surveys where applicable, which contain summary birth history (SBH) data. For Kenya, this includes a national MICS in 2000 and a 2011 MICS carried out in Nyanza province. For women in these datasets, MICS lists the name of the administrative unit within which they live. We also incorporate SBH data from IPUMS International census samples. For Kenya, this includes the 1989, 1999 and 2009 censuses. We cannot use full national censuses as they only provide information on urban/rural classification and not administrative unit. However, the IPUMS samples provide subnational classification by giving the name of an administrative area within which the woman lives. In both the MICS and IPUMS samples cases, the listed administrative unit is often smaller than our Admin-2 unit of interest. In these cases, we query the Google Maps API to get a GPS location for the administrative unit listed, assign that GPS coordinate to the appropriate Admin-2 unit.

2.2 Full birth history estimates

2.2.1 Discrete time survival model

Using the birth date and death date of each child in the DHS birth recode, we expand the child's data to contain an observation for each month of life. The response for each observation is whether the child died in that month. For a child that survives until its 5^{th} birthday, the child will contribute 60 observations to our dataset, all with zero responses. A neonatal death with contribute a single agemonth to our data with a response of 1. We use a design-based logistic regression model to fit the discrete hazards model via pseudolikelihood that accounts for the complex survey design of the DHS survey (Binder, 1983). We fit the following model using `svyglm` in the `survey` package in `R`:

$$\log\left(\frac{{}_1q_{x,a}^{its}}{1-{}_1\,q_{x,a}^{its}}\right) = \beta_a^{its}, \tag{1}$$

where i indexes the Admin-2 area, t represents 5-year periods 1980-1984,...,2010-2014, s indicates the survey and age x in age band a in months $[0,1), [1-12), [12-24), \dots, [48, 60)$. In this way, we fit a discrete hazards model for each area, time period and survey combination and allow the hazard to vary across the six age bands. When we fit this model for a specific area, we then use the delta method to produce a design-based estimate of the probability of death in an agemonth in age band a as

$${}_1\widehat{q}_{x,a}^{its} = \frac{\exp\left(\widehat{\beta}_a^{its}\right)}{1+\exp\left(\widehat{\beta}_a^{its}\right)}. \tag{2}$$

From this, we can construct a design-based estimate of U5MR as

$$\widehat{{}_{60}q_0}^{its} = 1 - \Pi_{a=1}^{6} \left(1 -_1 \widehat{q}_{x,a}^{its}\right)^{n_a}, \tag{3}$$

where $n_a = (1, 11, 12, 12, 12, 12)$ is the number of agemonths in each of the six agebands. Using the delta method and Equation 3, we also get an estimate of the variance of $\widehat{{}_{60}q_0}^{its}$, $\widehat{V}_{its}$ that accounts for the complex survey design.

2.2.2 Areas with zero deaths

Though there a few recent exceptions, most DHS are stratified at Admin-1 level crossed with urban/rural, not at Admin-2 level crossed with urban/rural. Sample size determinations for the survey are thus made at a coarser geographic granularity than that for which we would like to make estimates. As a result, some of the Admin-2 areas have only a few clusters and children in a given DHS survey to use in estimation. If there are no deaths in this small sample, we get a resulting designed-based estimate of ${}_{60}\widehat{q}_0^{its} = 0$ and, following our binomial model, a design-based estimate of $\widehat{V}_{its} = 0$. We know that in any given 5-year period, some children in an area will die regardless of how low the mortality rate is, so this is unbelievable. Additionally, this will create computational concerns in our smoothing model described in Section 2.4.

In these cases, we take all the data in areas i that neighbor the area with data problems. We use model-based methods to fit the following model to account for the complex survey design in `INLA`:

$$\begin{aligned}
y_{hc}^{its} | {}_{60}q_{0hc}^{its} &\sim \text{Bin}(n_{hc}^{its}, {}_{60}q_{0hc}^{its}), \\
\text{logit}({}_{60}q_{0hc}^{its}) &= \beta_h + \varepsilon_i + \varepsilon_c, \\
\varepsilon_i &\sim \text{N}(0, \sigma_1^2), \\
\varepsilon_c &\sim \text{N}(0, \sigma_2^2),
\end{aligned}$$

where i indexes area, c indexes cluster within area and h indexes the strata (urban or rurual) to which cluster c belongs. The fixed effect β_h accounts for stratification and the random effects account for dependencing within area and within clusters, respectively. We then get estimates of logit$(\widehat{{}_{60}q_0}^{its})$ and $\widehat{V}_{its}$ in the following way: We use these to get the following estimates for our small area of interest:

$$\begin{aligned}
P(\text{death}|\text{area } i) = {}_{60}q_0^{its} &\approx \sum_{h=1}^{H} \frac{N_{ih}}{N_i} \sum_{c=1}^{n_h} \frac{\hat{N}_{ihc}}{\hat{N}_{ih}} P(\text{death}|\text{strata } h, \text{ cluster } c, \text{ area } i) \\
&= \sum_{h=1}^{H} \frac{N_{ih}}{N_i} \sum_{c=1}^{n_h} \frac{\hat{N}_{ihc}}{\hat{N}_{ih}} {}_{60}q_{0hc}^{its}
\end{aligned} \tag{4}$$

where H is the number of strata in area i (always 2 for urban and rural), n_h is the number of clusters in strata h and area i. We calculate the number of people in an area and strata within area, N_i and N_{ih}, using gridded population density and urbanicity estimates. Figure 1 shows side by side comparisons of logit U5MR in all areas in Kenya before (L) and after (R) adjustment. There are three types of estiamtes to be adjusted on the left. Lamu, Embu, Laikipia, Trans-Nzoia, Kajiado and Samburu all have an estimate of $\widehat{{}_{60}q_0}^{i,80-84,2003} = 0$ and $\widehat{V}^{i,80-84,2003} = 0$. Marsabit and West Pokot have $\widehat{{}_{60}q_0}^{i,80-84,2003} \neq 0$ and $\widehat{V}^{i,80-84,2003} = 0$, which arises when all children every household (4 and 5 respectively) either survive until their 5^{th} birthday or die before it, we obviously do not trust this variance estimate when the sample size is so small. Lastly, Isiolo has no sampled clusters in the 2003 DHS, but we make no adjustments in this case. All adjusted estimates are pulled away from zero, and over all periods and surveys the mean CI width for adjusted areas is 2.01, compared to the mean CI width for unadjusted areas of 2.03. We do not recommend using this method for adjustment of a large proportion of areas; for Kenya, it is only 5% of area/time period/survey combinations.

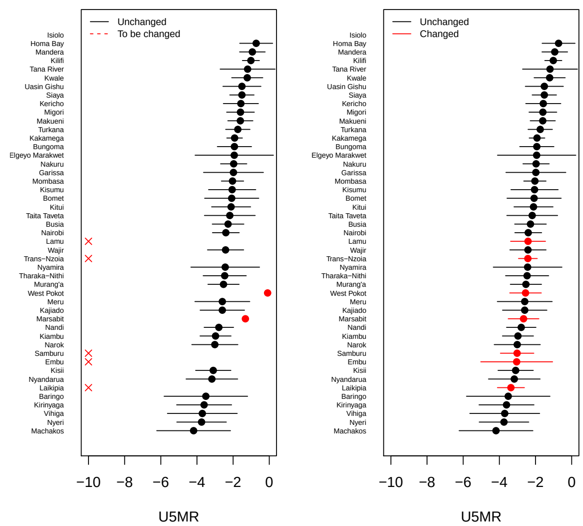

Figure 1: L: The direct estimates and CIs on the logit scale for data from the 2003 DHS for the period 1980-1985. Red x's are placeholders for areas with no deaths whose logit U5MR is undefined. R: The adjustment comparisons on the logit scale using additional information from neighboring areas.

2.3 Summary birth history estimates

The MICS and Census summary birth histories are analyzed using the Brass method (Trussell, 1975; Brass, 1968). This gives us a ${}_5\widehat{q}_0^{its}$ for a specific reference year t. We use a jackknife which deletes women to get an estimate of $\widehat{V}_{its}$.

2.4 Space-time smoothing

All estimates from the surveys and census can now be considered in the form $\left(y_{its} = \log\left(\frac{\widehat{{}_{60}q_0}^{its}}{1 - {}_5\widehat{q}_0^{its}}\right), \widehat{V}_{its}\right)$. Before smoothing we adjust for bias due to the HIV/AIDS epidemic using the methods described in Walter et al. (2012) (We use the following Bayesian hierarchical model to smooth:

$$
\begin{aligned}
y_{its} | \eta_{its}, \widehat{V}_{its} &\sim N(\eta_{its}, \widehat{V}_{its}) && (5) \\
\eta_{its} &= \mu + \nu_{s,SBH} + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it}, && (6) \\
& && (7)
\end{aligned}
$$

where μ is an overall intercept, $\nu_{s,SBH}$ is a fixed bias term for all data sources s providing SBH estimates, α_t and $\theta_i$are iid random effects for time, space, space-time interaction and survey, ϕ_i is an ICAR model for spatial smoothing and borrowing information between neighboring areas and γ_t is a random walk (RW2) model, and δ_{it} is a Type IV space-time interaction that accounts for dependency in space and time. All random effects involving time indices are fit on the yearly scale that aggregates up to 5-year periods. This methodological adjustment allows us to incorporate SBH data whose reference dates are estimated as single years with our FBH design-based estimates for coarser 5-year periods. This model is fit via `INLA` in `R`. We make estimates using all terms in the linear predictor aside from ν_s giving us smoothed estimates of ${}_{60}q_0^{it}$. This method does not work when ${}_{60}\widehat{q}_0^{its}$ is 0 as y_{its} is undefined, which motivates the need for our previously described adjustment method. Figure 2 gives smoothed results (L) for the 2010-2014 period and 95% PI widths (R). Figure 3 shows the initial estimates from various data sources in colors. The size of the dot representing the precision of the estimate, i.e. larger points should have greater influence in the smoothing. Black lines show the smoothed estimates and their 95% intervals.

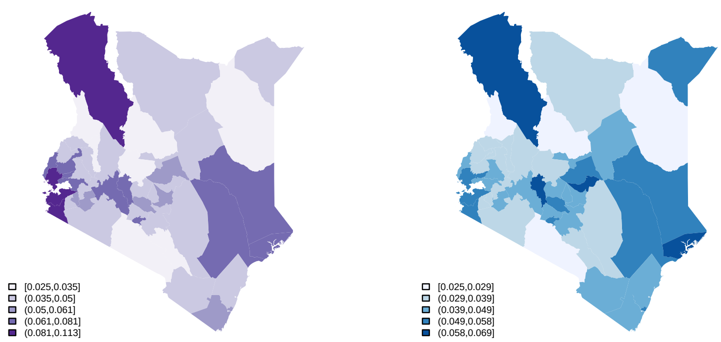

Figure 2: (L): Smoothed estimates for 2010-2014. (R): 95% PI widths for 2010-2014.

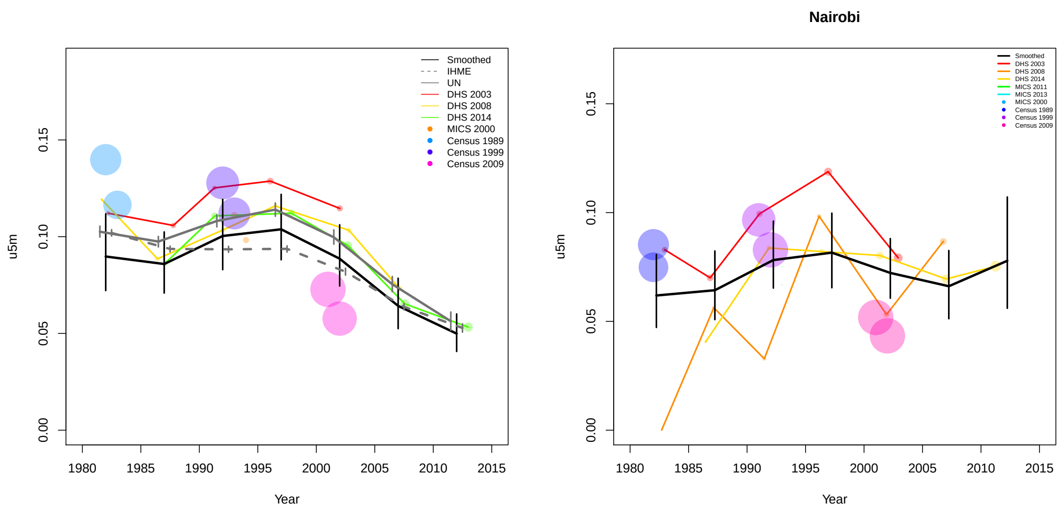

Figure 3: (L): National estimates for Kenya with comparisons to UN IGME and IHME. (R): Smoothed estimates for Nairobi across all time periods.