Urban Fraction for Stratified Beta-binomial Model

1 Summary

Stratified beta-binomial model produces estimates separately for urban and rural. An aggregation step is needed for obtaining overall estimates. Details of urban/rural stratification and entire procedure of aggregation could be found at section 3.4 and 3.5 in the original report. We recommend the reader to go over those sections before following the steps in this note. As summarized in the following equation, the beta-binomial model generates strata specific U5MR estimates and we need the urban/rural proportion for the under 5 population $q_{i,t}$ and $1 - q_{i,t}$ for all region i and time t to obtain the final overall estimates.

$$p_{i,t} = q_{i,t} \times U5MR_{i,t,R} + (1 - q_{i,t}) \times U5MR_{i,t,U}$$

The complete algorithm of finding those proportions.is implemented in two scripts: prepare_thresh.R, thresh.R and they should be run in the order listed. This vignette focuses on finding and downloading the data and hightlights some of the major steps.

2 Data Source

We need population density surfaces from worldpop (related scripts are prepare_thresh.R) and Admin-1 level urban population proportion (related scripts in thresh.R).

2.1 1km × 1km raster for whole population at the year of census

This raster is used for determining the pixel level urban/rural classification for the country of interest. To be consistent with the stratification, we want to use the surface at the year of the sampling frame construction. The year is usually the most recent census and it could be found in the DHS report. The population raster could be manually downloaded from https://www.worldpop.org/geodata/listing?id=75 (unconstrained individual countries 2000-2020 UN adjusted, 1km resolution). The name of the downloaded raster should be like 'xxx_ppp_2000_1km_Aggregated_UNadj.tif' where xxx is the three-letter abbreviation for the country (eg. zmb for Zambia). Then the .tif file should be manually placed into the folder 'Data\Zambia\Population'. Alternatively, the automated downloading scripts is:

```r
## pop.abbrev and pop_dir are prespecified in the scripts
setwd(pop_dir)

file <- paste0( pop.abbrev,'_ppp_',census.year,'_1km_Aggregated_UNadj.tif')

if(!file.exists(file)){

  url <- paste0("https://data.worldpop.org/GIS/Population/Global_2000_2020_1km_UNadj/",
                census.year, "/", toupper(pop.abbrev),"/",
                pop.abbrev,'_ppp_',census.year,'_1km_Aggregated_UNadj.tif')

  download.file(url, file, method = "libcurl",mode="wb")
}
```

2.2 100m × 100m raster for under-5 population at the years for estimation

These rasters are used for aggregating strata specific U5MR. We use worldpop population density broken down by age and sex (available at https://www.worldpop.org/geodata/listing?id=30). For each year, four rasters are needed: 0-1 male (xxx_m_0_year.tif), 1-5 male (xxx_m_1_year.tif), 0-1 female (xxx_f_0_year.tif) and 1-5 female (xxx_f_1_year.tif), where xxx is the three-letter abbreviation for the country. For estiamtes in the past 9 years, 36 rasters in total should be donwloaded. Because of website constraint, the automated downloading scripts might not work well. In that case, manually downloaded .tif files should be put into the folder 'Data\Zambia\Population'. The automated downloading scripts is:

```r
#! downloading might take a long time, especially for big countries

pop.year <- beg.year:end.year  ## population surface year

setwd(pop_dir)

options(timeout = 1000) # adjust this time, should be longer than each download
for(year in pop.year){
  print(year)
  for(age in c(0, 1)){
    for(sex in c("f", "m")){
      file <- paste0(pop.abbrev,'_', sex, '_', age, '_', year,'.tif')
      if(!file.exists(file)){
        url <- paste0("https://data.worldpop.org/GIS/AgeSex_structures/Global_2000_2020/",
                      year, "/", toupper(pop.abbrev), "/", pop.abbrev, "_",
                      sex, "_", age, "_", year, ".tif")
        download.file(url, file, method = "libcurl",mode="wb")
      }
    }
  }
}
```

2.3 Admin-1 level urban population fraction

The thresholding algorithm requires knowledge of urban fraction at admin-1 level. Such information could usually obtained from the DHS reports, or census summary tables. We aim to prepare a cleaned table with one column indicating the Admin-1 region name and another column the urban population fraction. Note that this step involves country specific treatment and needs extra consideration. The general search process is summarized below and the options are ordered by their priority. We illustrate the process using Zambia 2018 as an example.

1. We recommend the reader to first check out the DHS report of that specific survey. Ideally, the sampling frame information will appear in appendix A. However, in our example, the admin-1 level urban population fraction is not available, and we only have household urban fraction (report available at https://dhsprogram.com/pubs/pdf/FR361/FR361.pdf).

2. The second option is to check out the report for an earlier DHS survey for the same country. DHS surveys usually use the same sampling frame as the most recent census. As the census is less frequent than DHS surveys, it is likely (but not necessarily) that multiple DHS surveys share the same sampling frame. One should be cautious about using this approach as the sampling frame might be updated between surveys (such as in Nigeria). In our example, we will go through the final report for DHS 2013-2014 in Zambia (available at https://dhsprogram.com/pubs/pdf/FR304/FR304.pdf). This time, we found the information in table A.1.

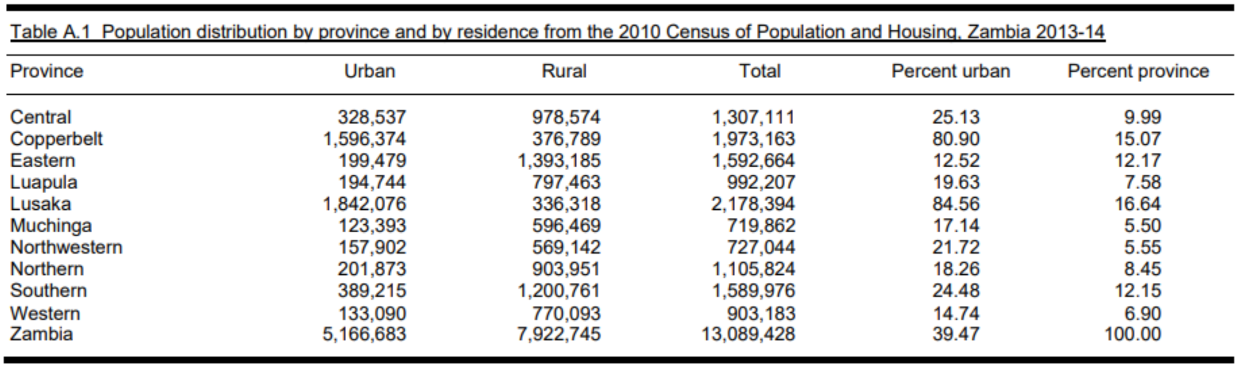

Table A.1 Population distribution by province and by residence from the 2010 Census of Population and Housing, Zambia 2013-14

Province	Urban	Rural	Total	Percent urban	Percent province
Central	328,537	978,574	1,307,111	25.13	9.99
Copperbelt	1,596,374	376,789	1,973,163	80.90	15.07
Eastern	199,479	1,393,185	1,592,664	12.52	12.17
Luapula	194,744	797,463	992,207	19.63	7.58
Lusaka	1,842,076	336,318	2,178,394	84.56	16.64
Muchinga	123,393	596,469	719,862	17.14	5.50
Northwestern	157,902	569,142	727,044	21.72	5.55
Northern	201,873	903,951	1,105,824	18.26	8.45
Southern	389,215	1,200,761	1,589,976	24.48	12.15
Western	133,090	770,093	903,183	14.74	6.90
Zambia	5,166,683	7,922,745	13,089,428	39.47	100.00

Figure 1: Urban fraction found in Zambia DHS 2013-2014 report

3. If the DHS survey uses the same sampling frame as a census, summary report for that census might contain Admin-1 urban fraction. From table 2.3 in the summary report for Zambia census 2010, we obtain the same information.

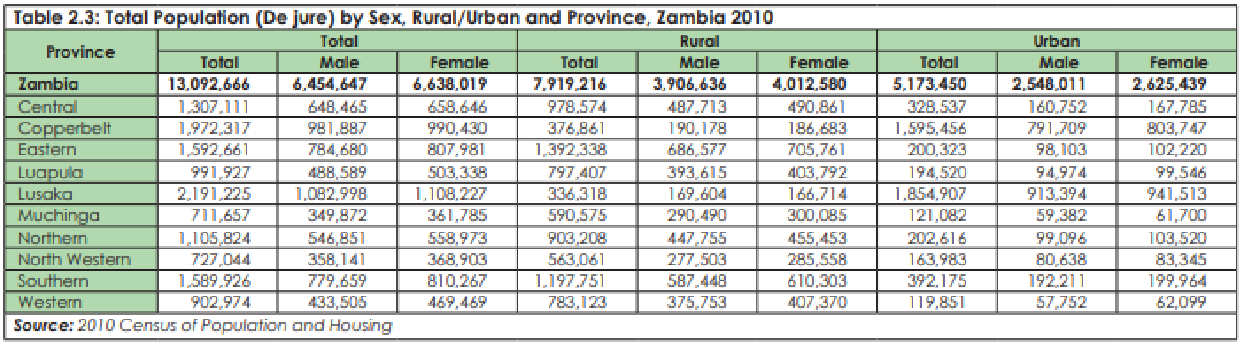

Table 2.3: Total Population (De jure) by Sex, Rural/Urban and Province, Zambia 2010

Province	Total			Rural			Urban		
	Total	Male	Female	Total	Male	Female	Total	Male	Female
Zambia	**13,092,666**	**6,454,647**	**6,638,019**	**7,919,216**	**3,906,636**	**4,012,580**	**5,173,450**	**2,548,011**	**2,625,439**
Central	1,307,111	648,465	658,646	978,574	487,713	490,861	328,537	160,752	167,785
Copperbelt	1,972,317	981,887	990,430	376,861	190,178	186,683	1,595,456	791,709	803,747
Eastern	1,592,661	784,680	807,981	1,392,338	686,577	705,761	200,323	98,103	102,220
Luapula	991,927	488,589	503,338	797,407	393,615	403,792	194,520	94,974	99,546
Lusaka	2,191,225	1,082,998	1,108,227	336,318	169,604	166,714	1,854,907	913,394	941,513
Muchinga	711,657	349,872	361,785	590,575	290,490	300,085	121,082	59,382	61,700
Northern	1,105,824	546,851	558,973	903,208	447,755	455,453	202,616	99,096	103,520
North Western	727,044	358,141	368,903	563,061	277,503	285,558	163,983	80,638	83,345
Southern	1,589,926	779,659	810,267	1,197,751	587,448	610,303	392,175	192,211	199,964
Western	902,974	433,505	469,469	783,123	375,753	407,370	119,851	57,752	62,099

Source: *2010 Census of Population and Housing*

Figure 2: Urban fraction found in Zambia census 2010 report

4. The last option is to weight the number of households in urban/rural by urban/rural specific average household sizes. This approach is not ideal because the average household sizes are usually only available at national level and rounding will further compromise the accuracy. In the Zambia 2018 example, we might find information from Table A.1 and Figure 2.9 in the DHS Zambia 2018 report.

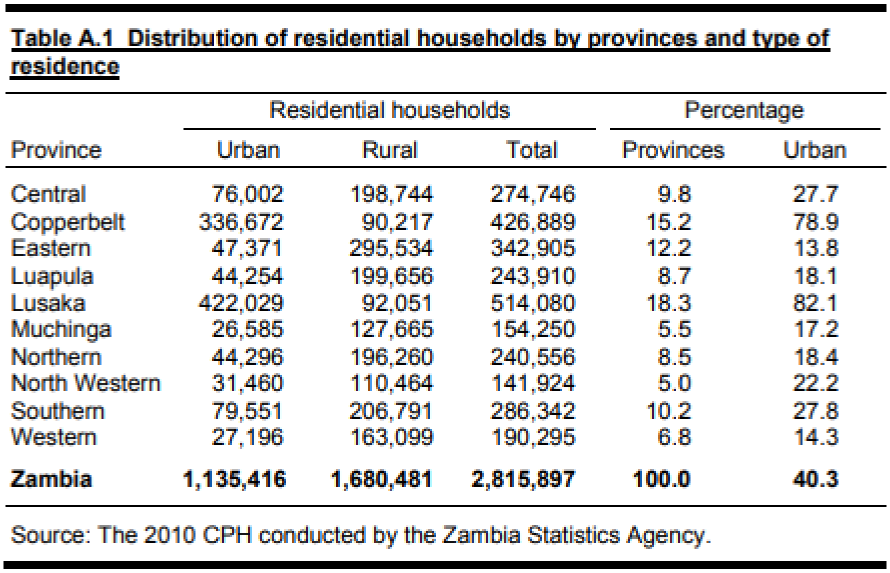

Table A.1 Distribution of residential households by provinces and type of residence

	Residential households			Percentage	
Province	Urban	Rural	Total	Provinces	Urban
Central	76,002	198,744	274,746	9.8	27.7
Copperbelt	336,672	90,217	426,889	15.2	78.9
Eastern	47,371	295,534	342,905	12.2	13.8
Luapula	44,254	199,656	243,910	8.7	18.1
Lusaka	422,029	92,051	514,080	18.3	82.1
Muchinga	26,585	127,665	154,250	5.5	17.2
Northern	44,296	196,260	240,556	8.5	18.4
North Western	31,460	110,464	141,924	5.0	22.2
Southern	79,551	206,791	286,342	10.2	27.8
Western	27,196	163,099	190,295	6.8	14.3
Zambia	**1,135,416**	**1,680,481**	**2,815,897**	**100.0**	**40.3**

Source: The 2010 CPH conducted by the Zambia Statistics Agency.

Figure 3: Urban fraction for households found in Zambia census 2018 report

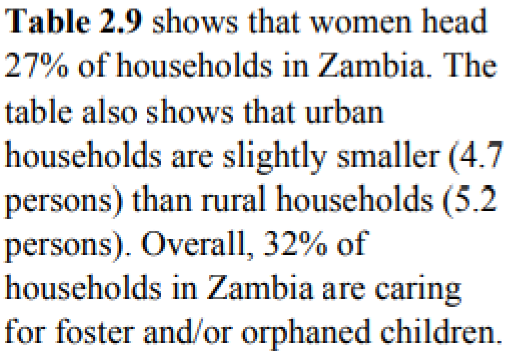

Table 2.9 shows that women head 27% of households in Zambia. The table also shows that urban households are slightly smaller (4.7 persons) than rural households (5.2 persons). Overall, 32% of households in Zambia are caring for foster and/or orphaned children.

Figure 4: Average households size found in Zambia census 2018 report

After the Admin-1 urban population fractions are located, we will conduct ad hoc data cleaning to prepare a data frame. We recommend first copy the table from report into a .txt file and clean it in R. The .txt file should be placed under '\Data\country\'. We show some details of this step using Zambia as an example. Note that additional scripts might be needed to accommodate country specific situation.

Generally, population counts in the table will be displayed with comma. The following scripts get rid of those commas.

```
frame<-read.delim(paste(country.abbrev, "frame_urb_prop.txt", sep = "_"),
                  header = FALSE,sep=' ')

# identify column for fraction (need additional processing in general)
frame[,c(2,4)] <- lapply(frame[,c(2,4)],   ## function to remove comma in numbers
                         function(x){as.numeric(gsub(",", "", x))})
frame$frac <- frame$V4/frame$V2
```

Another difficulty is to match the Admin-1 regions names from the table and from the GADM shapefile. A greedy algorithm might be adopted to automate the matching process. It could also be done manually.

```
# greedy algorithm to match admin names
adm1.ref <- expand.grid(tolower(frame$V1),
                        tolower(admin1.names$GADM)) # Distance matrix in long form
names(adm1.ref) <- c("frame_name","gadm_name")
### string distance,  jw=jaro winkler distance, try 'dl' if not working
adm1.ref$dist <- stringdist(adm1.ref$frame_name,
                            adm1.ref$gadm_name, method="jw")

greedyAssign <- function(a,b,d){
  x <- numeric(length(a)) # assgn variable: 0 for unassigned but assignable,
  # 1 for already assigned, -1 for unassigned and unassignable
  while(any(x==0)){
    min_d <- min(d[x==0]) # identify closest pair, arbitrarily selecting 1st if multiple pairs
    a_sel <- a[d==min_d & x==0][1]
    b_sel <- b[d==min_d & a == a_sel & x==0][1]
    x[a==a_sel & b == b_sel] <- 1
    x[x==0 & (a==a_sel|b==b_sel)] <- -1
  }
  cbind(a=a[x==1],b=b[x==1],d=d[x==1])
}
```

```
match_order<-data.frame(greedyAssign(adm1.ref$frame_name,
                                     adm1.ref$gadm_name,
                                     adm1.ref$dist))
```

Finally, a cleaned reference table will be created with one column containing Admin-1 region names consistent with GDAM and another column containing urban population fraction. We urge the user to check the resulting data set for mismatch.

```
# create reference table
ref.tab <- admin1.names
ref.tab$matched_name <- frame$V1[match_order$a] ### check!!!
ref.tab$urb_frac <- frame$frac[match_order$a]
```

3 Checking Accuracy of Thresholding

The thresholding algorithm will yield a pixel level classification map for urban rural status. As a sanity check, we could use this map to assign the sampled clusters to urban/rural, and then compare with the classification (stratification) used in the survey.

In DHS surveys, the true precise locations for clusters are not given for confidentiality. Instead, we only observe the jittered version of the locations. This is potentially disruptive to our classification results as the urban cluster might be moved to a rural area. There is a section called 'Correct urban cluster' in the scripts prepare_thresh.R, where we move urban clusters to more populated pixel to alleviate the side effect of jittering.

The user could also choose to validate using uncorrected locations, though some modifications for the scripts are required.