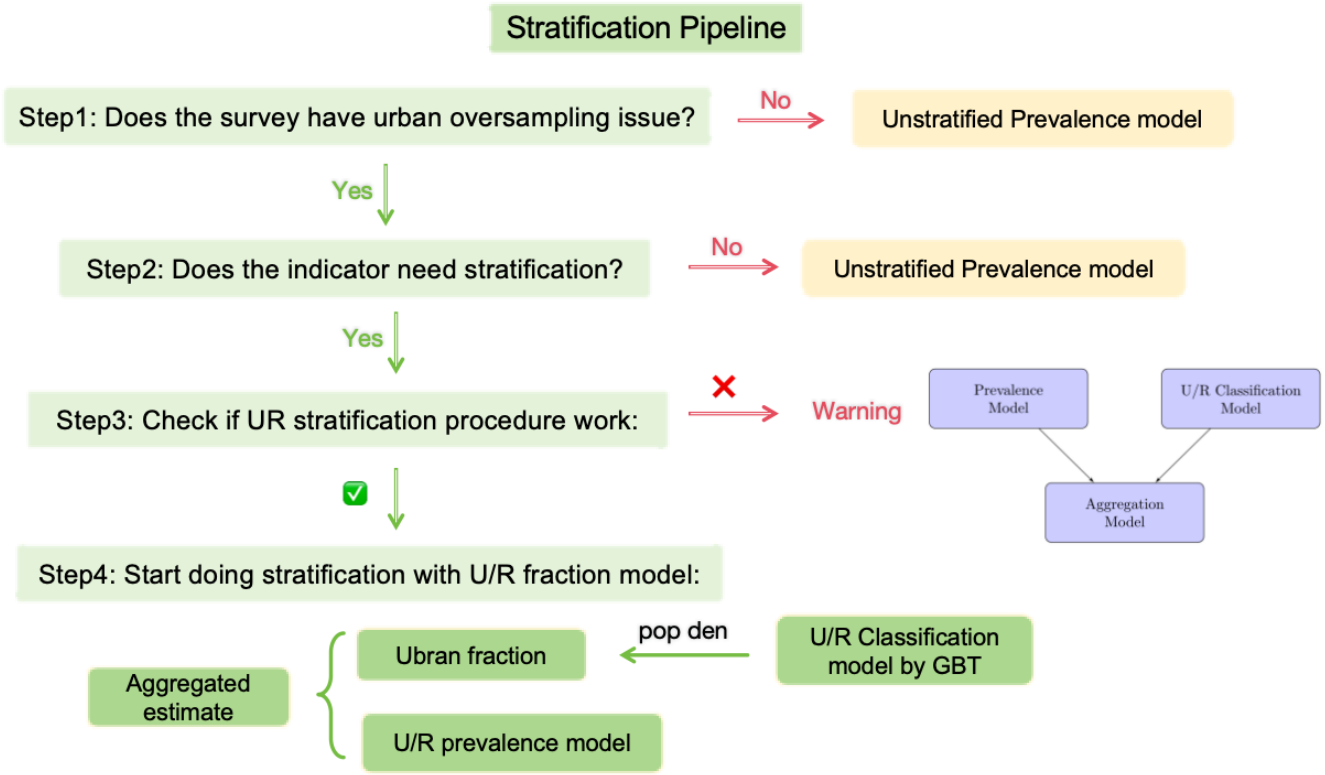
Stratification Pipeline

Jitong Jiang, Yunhan Wu

Stratification model overview



Unstratified Prevalence model

$$Y_c \mid p_c, d \sim \text{BetaBinomial}(n_c, p_c, d),$$

$$p_c = \text{expit}(\alpha + e_{i[s_c]} + S_{i[s_c]}).$$

$$b_{i[s_c]} = e_{i[s_c]} + S_{i[s_c]}, b \sim \text{BYM2}(\sigma_s^2, \phi)$$

Step 1: Determine oversampling issue:

Chi-square test

Observed test statistic :

$$\chi^2 = \sum_{i=1}^{M} \chi_i^2 = \sum_{i=1}^{M} \frac{(O_{u_i} - E_{u_i})^2}{E_{u_i}} + \frac{(O_{r_i} - E_{r_i})^2}{E_{r_i}} \qquad (1)$$

asymptotically follows $\chi^2(M)$ if there is no low count

Where:

- O_{u_i} is the observed number of urban EA counts in region i,
- O_{r_i} is the observed number of rural EA counts in region i,
- $E_{u_i} = N_i \times p_i^u$ is the expected number of urban EA counts in region i, where N_i is the total sample size for region i and p_i^u is the urban fraction in sampling frame,
- $E_{r_i} = N_i - E_{u_i}$ is the expected number of rural EA counts in region i.

Step 1: Determine oversampling issue

Monte Carlo simulation:

$$O_{u_i}^{(sim)} \sim \text{Binomial}(N_i, p_i^u) \tag{2}$$

$$O_{r_i}^{(sim)} = N_i - O_{u_i}^{(sim)} \tag{3}$$

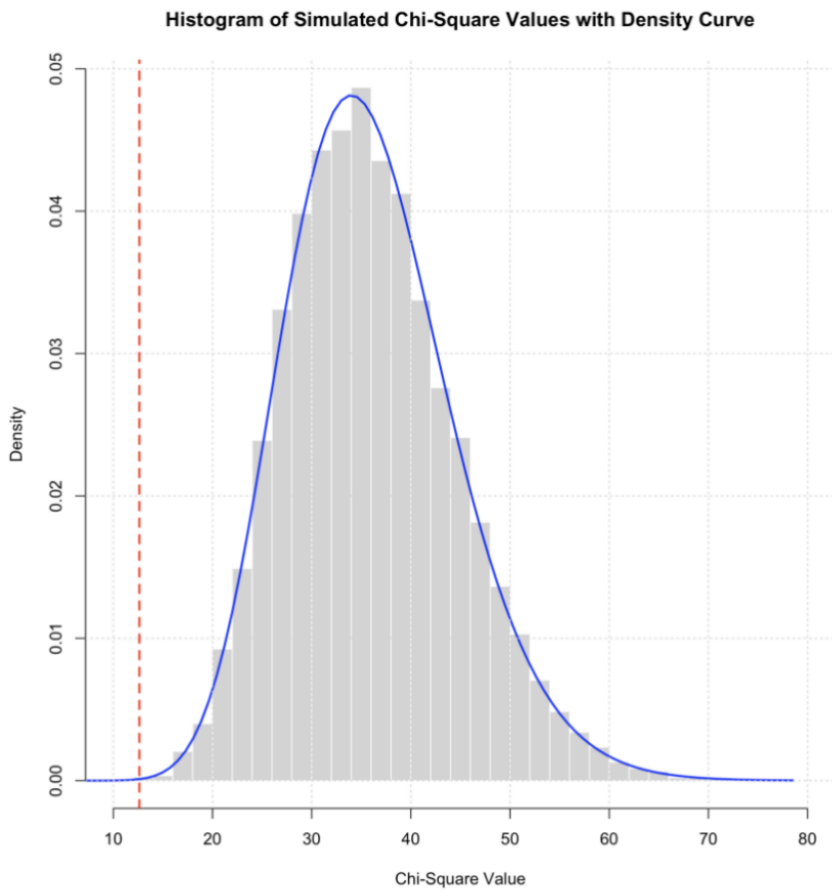
For each simulation j, the chi-square statistic $\chi^2_{sim_j}$ is calculated similarly to the observed chi-square statistic.
We simulate 1000 times, i.e., $j = 1 : 1000$.
The p-value is:

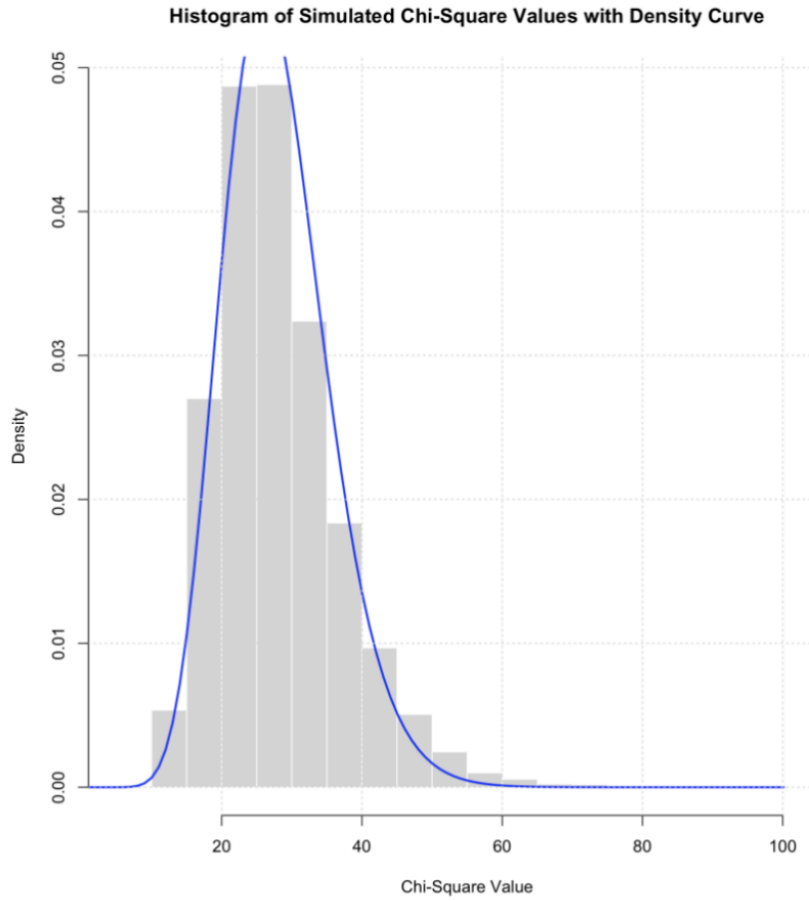
$$p\text{-value} = \frac{\sum_{j=1}^{1000} \mathbb{I}(\chi^2_{sim_j} \geq \chi^2_{obs})}{1000} \tag{4}$$

Step 1: Determine oversampling issue:

- Nigeria: Observed Chi-square = 12.62625, p-value = 0.9996
- Malawi: Observed Chi-square = 331.7202, p-value = 0

(Left: Nigeria, Right: Malawi)





Step 2: Determine stratification or not

For each indicator, we decide the need to do stratification based on a Likelihood Ratio Test:
Model1 vs Model2, where

- Model1: value $\sim$ admin1
- Model2: value $\sim$ admin1 + UR

Step 3: Check UR stratitication procedure

Goal: compare stratified model vs unstratified model against direct estimates

Prevalence model: admin1 fixed effect, no smoothing.

- Unstratified prevalence model:

$$Y_c \mid p_c, d \sim \text{Beta-Binomial}(n_c, p_c, d),$$

$$p_c = \text{expit}(\alpha_{i[s_c]}).$$

- Stratified prevalence model:

$$Y_c \mid p_c, d \sim \text{Beta-Binomial}(n_c, p_c, d),$$

$$p_c = \text{expit}(\alpha_{i[s_c]} + \gamma \times I(s_c \in \text{urban})).$$

Step 3: Check UR stratification procedure

- Aggregated stratified prevalence model at admin1:

$$p_i^S = p_{i,U} \times q_i + p_{i,R} \times (1 - q_i)$$

- Unstratified prevalence model at admin1:

$$p_i^{US} = p_{i[s_c]}$$

- Direct estimate model at admin1:

$$p_i^D = \frac{\sum_{j \in S_i} y_j w_j}{\sum_{j \in S_i} w_j}$$

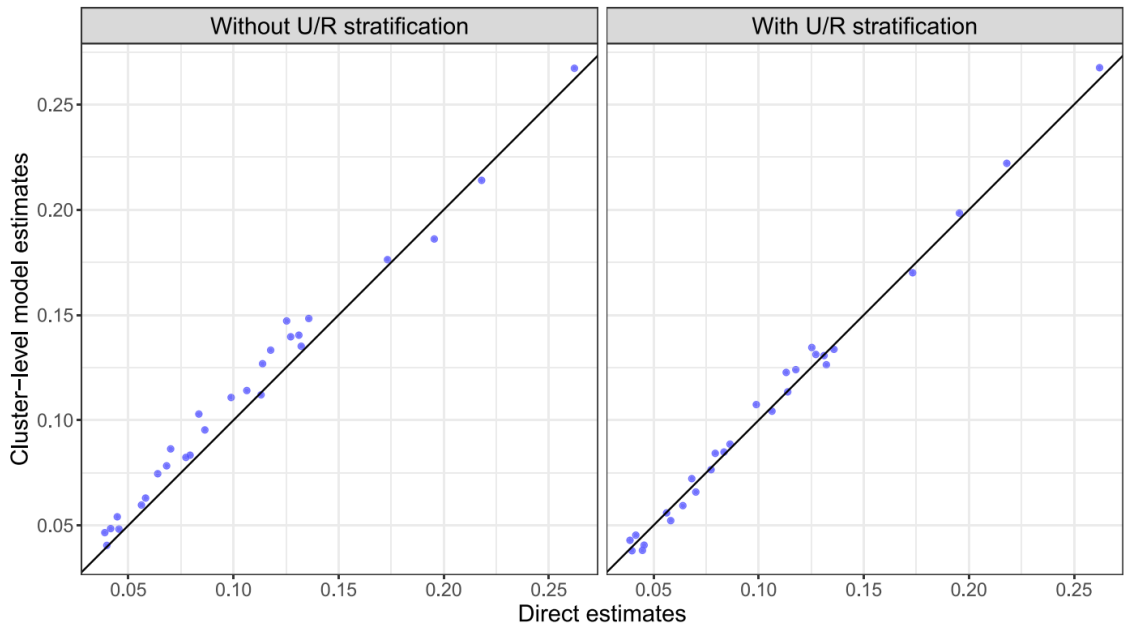
Step3: Check UR stratification procedure

Comparison: Compare stratified and unstratified models to determine which is closer to the direct estimate.

$$\text{Weighted absolute bias}(\hat{p}, p^D) = \frac{1}{\sum_{i=1}^{M} N_i} \sum_{i=1}^{M} N_i \times |\hat{p}_i - p_i^D|$$

N_i: population of admin region i

$\hat{p} : p^{US}$ or p^S



Step 4: Stratification for admin2/3

Prevalence model for admin2/3:

$$Y_c \mid p_c, d \sim \text{BetaBinomial}(n_c, p_c, d),$$

$$p_c = \text{expit}(\alpha + \gamma \times I(s_c \in \text{urban}) + e_{i[s_c]} + S_{i[s_c]}).$$

$$b_{i[s_c]} = e_{i[s_c]} + S_{i[s_c]}, b \sim \text{BYM2}(\sigma_s^2, \phi)$$

U/R fraction: q_i
Aggregation Estimate:

$$p_i^S = p_{i,U} \times q_i + p_{i,R} \times (1 - q_i)$$

U/R Classification model: Covariates & Jittering

To get the U/R fraction q_i, we need to have a U/R indicator map for each country.

- Model: A classification model using Gradient Boosted Trees (GBT)
- Training set: For each DHS cluster with U/R classification (jittering the coordinate), include population density (WorldPop), nighttime light (NOAA, VIIRS), and administrative region (as U/R classification may vary by region),
- Prediction set: each national pixel with information of population density, nighttime light, and administrative region.
- Outcome: the probability of being urban for each national pixel.
- Calibration: since the predicted probabilities obtained directly from the classification model are not well calibrated, calibrate it by the sampling frame.

U/R Predicated Indicator Surface

Deriving an U/R indicator surface using population thresholds: Urban classification is based on population density, ranked by predicted urban probability. Pixels are classified as urban once their cumulative population fraction reaches the regional threshold; the rest are rural.
Then we obtain the fraction by aggregating with the subpopulation:

$$q_j = \frac{\sum_{g \in A_j^*} I(g^* \in urban) \times N_g^*}{\sum_{g \in A_j^*} N_g^*}$$

where

N_g^* : indicator's corresponding population at pixel g.

A_j^* : target region