Estimating populations at high resolution in data-poor regions using commonly available population-weighted household survey data

Douglas R. Leasure[a,1], **Claire A. Dooley**[a], **Gianluca Boo**[a], **Édith C. Darin**[a], **and Andrew J. Tatem**[a]

[a]WorldPop, Geography and Environmental Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

This manuscript was compiled on July 10, 2019

250 words.

demography | international development | Bayesian statistics | household surveys

Population estimates that are accurate and up-to-date are critical for government planning and development projects in low and middle income countries where recent census results may not be available and field surveys designed to collect survey data specifically for population estimation can be logistically challenging. Household surveys are routinely conducted in these countries, often with national coverage, and they generally enumerate all people in households where surveys are conducted. There are two challenges preventing these data from being used to estimate population sizes at high resolution: 1) access to sensitive household survey results are protected due to privacy concerns and 2) household surveys generally select sampling locations using a population-weighted sampling scheme to conduct more surveys in high-density urban areas that would be under-represented in a random sample.

Our objectives are:

1. Demonstrate a Bayesian weighted-likelihood model in a simulation environment to correct bias inherent in a weighted sample.

2. Estimate population sizes in surveyed areas of Kinshasa province in the Democratic Republic of Congo using a random sample collected in the field.

3. Apply the weighted-likelihood approach to estimate population sizes in surveyed areas of Kinshasa using an independent population-weighted sample collected in the field.

We developed a Bayesian weighted-likelihood model for use with standard household survey data and assessed its performance in simulated and real-world environments. The population estimates produced here for surveyed areas in Kinshasa province are not intended for uses beyond demonstrating the method.

Results.

Discussion.

Materials and Methods

A. Simulated Data. Populations were simulated using 10,000 random draws from log-normal distributions to represent the number of people in 10,000 different locations. An urban settlement type was simulated using a median of 750 and a standard deviation of 500 to parameterise the log-normal distribution. A rural settlement type was simulated using a median of 100 and a standard deviation of 250. In simulations that included both settlement types, half of the locations were urban and half were rural.

Simulated populations were sampled using either random sampling, population-weighted sampling, or both. A total of 1,000 samples were collected for each simulation. In simulations that included both types of sampling, half of the samples were random and half were population-weighted. For simulated populations that included two settlement types, stratified sampling was implemented to ensure that an equal number of samples was collected from each settlement type. Population-weighted sampling used sample weights w_j to define the probabilities for each location j being sampled:

$$w_j = \frac{N_j}{\sum_{j=1}^{J} N_j}$$

where N_j is the population size at location j, and J is the total number of simulated locations (*i.e.* 10,000).

Six scenarios were simulated:

1. Random sampling from urban locations
2. Random sampling from urban and rural locations
3. Weighted sampling from urban locations
4. Weighted sampling from urban and rural locations
5. Random and weighted sampling from urban locations
6. Random and weighted sampling from urban and rural locations

B. Real Data. Microcensus surveys were conducted throughout Kinshasa province in Democratic Republic of the Congo in 2017 and 2018 for the purpose of estimating and mapping the population. Microcensus surveys enumerated all people within a sample of locations that each contained about 3 hectares of settled area within a clearly defined survey boundary (i.e. a survey cluster). The 2017 microcensus collected data from 515 locations using a stratified random sampling design, and the 2018 microcensus collected data from 411 locations using a stratified population-weighted sampling design.

For the 2017 microcensus, stratification of samples was based on settlement types from the LandScanHD database () which included urban, hamlet, and rural settlement types. For the 2018 microcensus, stratification of samples was based on k-means clustering of principal components derived from a set of XXX covariates (100 m grid cells)

Significance Statement

Leasure, Dooley: methods development, simulations, data analysis; Boo, Darin: field data collection/prep, data analysis; Tatem: project planning and implementation. All authors contributed to writing.

No conflicts of interest.

[1]Corresponding author. E-mail: doug.leasure@gmail.com

from the WorldPop Global project (). This classified each 100 m grid cell as one of three settlement types.

A one-stage survey design was used in which a household listing for each survey cluster was created on the same day as household surveys were conducted. Household surveys produced a roster of individuals in the household with basic information about them including age, sex, education, etc. If no respondent was not available from a household, a neighbor was asked to answer questions on their behalf. For households where no respondent could be identified, household sizes were imputed using the mean household size for the cluster.

C. Data Analysis. We used a log-normal weighted-likelihood model to represent the distribution of population sizes among locations:

$$y_i \sim LogNormal(log(\mu_t), \tau_{t,i})$$

where y_i is the observed number of people at sampled location i, and mu_t is the median population size for the settlement type t to which location i belongs. Sample weights w_i were used to calculate model weights v_i to account for the sampling bias:

$$v_i = \frac{w_i^{-1}}{\sum_{i=1}^{I} w_i^{-1}}$$

where i is a sampled location and I is the total number of sampled locations (*i.e.* 1,000). Model weights are the inverse of sample weights that are re-scaled to sum to one among all sampled locations.

τ_i is a location-specific estimate of precision (*i.e.* the inverse of variance) that is dependent on the model weights v_i and a global estimate of precision $\bar{\tau}_t$:

$$\tau_{t,i} = \bar{\tau}_t v_i$$

We defined the prior distributions of μ_t and $\bar{\tau}_t$ using uninformative uniform distributions:

$$\mu_t \sim Uniform(0, X)$$

$$\bar{\tau}_t \sim Uniform(0, X)$$

$\bar{\tau}_t$ cannot be used to predict y in new locations where model weights are not available. So, we used a weighted average of $\tau_{t,i}$ to derive an estimate of precision for each settlement type that can be used for predictions in new locations:

$$\theta_t = \left(\frac{\sum_{i=1}^{I_t} v_i \sqrt{\tau_{t,i}^{-1}}}{\sum_{i=1}^{I_t} v_i}\right)^{-2}$$

where I_t is the total number of samples from settlement type t. The weighted average was calculated based on the standard deviation $\sqrt{\tau_{t,i}^{-1}}$ rather than precision $\tau_{t,i}$ because ?????. This derived estimate of precision was used for making posterior predictions without the need for model weights:

$$\hat{y}_t \sim LogNormal(log(\mu_t), \theta_t)$$

ACKNOWLEDGMENTS. This work was part of the GRID^3 project (Geo-Referenced Infrastructure and Demographic Data for Development) funded by the Bill and Melinda Gates Foundation and the United Kingdom Department of International Development. The project is a collaboration between the WorldPop Research Group at the University of Southampton, the Flowminder Foundation, the United Nations Population Fund (UNFPA), and the Center for International Earth Science Information Network (CIESIN) within the Earth Institute at Columbia University. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.