Representative sampling off the rural road: Lessons from a sampling design and survey implementation effort in Rwanda

Stuart H. Sweeney[1], Jessica Marter-Kenyon[2], and Sophia D. Arabadjis[3]

[1]*Department of Geography, University of California Santa Barbara*
[2]*Innovation Lab for Peanut, University of Georgia*
[3]*Institute for Implementation Science and Population Health, City University of New York*

Abstract

Demographic surveys implemented in rural Africa often proceed without access to a reliable sampling frame. In the absence of such a frame, a rule of thumb is to select every x^{th} house along a road. This may result in non-representative sampling and biased inference for several reasons. This paper reports on a novel survey sampling approach implemented in Rwanda and the associated construction of sampling weights. The intent of the survey is evaluate the impacts of Rwanda's villagization policy on aspects of family formation and time to first birth. We develop a sampling approach that implements a spatial inhibition process combined with aggregate data on certain areal characteristics of Rwandan districts and sectors. After reviewing the sampling algorithm, we describe the resampling approach used to estimate own and cross probabilities of inclusion, which are then used in a Horvitz-Thompson estimator. We evaluate the bias and efficiency of our approach using simulated data and relative to other large-scale survey information (census and Demographic and Health Survey). The paper closes with a description of lessons learned during the implementation phase of the survey.

Keywords

Survey sampling without replacement, Horvitz-Thompson estimator, family formation, fertility, Villagization policy, Rwanda

Background

Rwanda is an east African country with stunning terrain characterized by mountains in the north and west, transitioning to a flat plateau in the east. The varied environments contain high biodiversity which are protected by several reserve areas, and span 12 agroecological zones with an associated range of livelihoods. It is one of the most densely populated countries in Africa. The settlement structure includes a capital city, several secondary cities, about 15,000 villages (groupings of houses) of varying sizes and spatial organization. Some are highly structured, others more organic, and there are also isolated houses scattered between villages.

For more than thirty years, it has been the stated intention of the Rwandan government to group approximately 90% of its rural population into small, clustered villages (or *imidigudu*). Villagization is a lynchpin in Rwanda's plan to achieve its environment and development vision. It is viewed as the critical precursor for facilitating the rest of the agrarian transformation. It frees up land for commercial production, improves the legibility of farmers and the government's ability to encourage and monitor compliance with land use policies, and pushes smallholder farmers into contract farming and off-farm jobs. Today, the *imidugudu* policy is 30 years old, and more than 7 million rural people are living in grouped settlements.

We undertook our survey data collection to understand how the villagization processes was impacting household formation and other demographic outcomes. Specifically, the villagization policy was thought to increase the costs associated with household formation and our expectation was that age and first marriage and the duration to first birth, would increase for couples living in formal "villagized" settlements relative to those in informal settlements. At the same time, formal villages are expected to have higher connectivity to services, such as health clinics and schools, which may create higher access to contraception, for example.

The central problem for our sampling was that no government household registry exist that spans villagized (imidugugu developments) and non-villagized 'traditional' settlments. We needed to ensure a sufficiently large sample size in each group that it would support our inferential comparison. It was also important- as it is in all survey research- to introduce as much randomization as possible when selecting farm households to interview in each settlement type.

Both of these criteria are rather difficult to achieve in the Rwandan rural context, as is the case in many other African and developing countries [3, 1]. For one thing, population/household registers either do not exist or are difficult to access. Additionally, relying solely on enumerators to select interview households can lead to a skewed sample, as enumerators may not be able to identify settlement types visually and they are likely to choose households that are relatively easy to access.

The total sample size, number of districts and cells within districts sampled, and number of households per village or nonvillage points were not based on power calculations. Instead those decisions were based on trying to both maximize the sample size while maintaining variation in areas sampled but also with an eye towards efficiency in data collection by the survey team.

The research was approved by the Institutional Review Board of the University of California at Santa Barbara (UCSB). Additionally, research permits were granted by the Ministry of Education of the Government of Rwanda with the University of Rwanda serving as the sponsoring institution. Visits were made to each district office to inform local administrators of the research and to obtain their support.

Methods

Sampling Algorithm

The official territorial organization of Rwanda includes four provinces (plus the city of Kigali), 30 districts, 416 sectors, 2,148 cells and 14,837 sub-cells (villages, or imidugudu, which contain both planned villages and isolated settlements). A registry of imidugudu developments, including their spatial centerpoint, exists but there is no associated registry of informal villages. The sampling algorithm proceeded according to the following steps.

1. Districts: We selected four districts in the country, one in each of the four provinces: Gatsibo (East), Nyanza (South), Nyamasheke (West) and Musanze (North). Four were chosen because of resource limitations. The specific districts were chosen because, together, they represent ten of the twelve Rwandan agroecological zones and therefore provide us a representative cross-section of the impacts of villagization on various outcomes of interest under different livelihood settings, and increase the amount of variability in the survey sample.

2. Cells: Within each district we clustered our sample at the cell level, adding adjacent cells until there were at least 5 formal villages (imidugugu developments) within the cell or cell cluster. GIS data on administrative boundaries and formal village centerpoints was obtained from the Rwanda Housing Authority. In Nyanza and Gatsibo, we had 14 cells each. In Nyamasheke and Musanze we had 11 and 12 cell clusters, respectively. Selected cells represented between 20% and 49% of the total cells in a district, depending on the district.

3. Formal Villages: Within each cell/ cell group, we randomly selected five planned village sites. Village sites were chosen to select dispersed, non-adjacent locations.

4. Informal Villages: We then used an inhibition spatial point process to seed a search location for informal villages. The informal village seed locations were generated sequentially such that a corresponding seed point was located, at minimum, beyond a threshold distance from formal villages and any other informal village seed points already generated. The seeds were then used by the field team to initiate a search for the closest informal village. In practice this was accomplished with the aid of Google earth imagery and then implemented in the field. [Note: Reduced the potential for bias in selecting the isolated settlement location starting points (Kondo et al. 2014).]

5. Visual Manipulation: Once the locations were chosen we imported them to Google Earth, which has sufficient image resolution to be able to visually detect settled areas and the extent of their relative grouping or isolation. It also has sufficient resolution to see the roofs of individual households. The imagery was up-to-date within one or two months of the sampling.

   We visually scanned the satellite imagery; if a selected non-village location turned out to be in a forest or field, or otherwise far away from a house, it was manually moved it to the nearest house. We could have done this in a more rigorous fashion, for example by quantitatively determining the distance between the original point and all nearby houses, then selecting the closest house. Between 88% (Nyamasheke) and 95.4% (Gatsibo) of non-village points were manually manipulated with the aid of Google Earth.

6. Households: Once the formal and informal village locations were finalized, we downloaded their coordinate locations into GPS devices. Enumerators were each given a GPS device

(Garmin GPSMAP 64st Handheld) containing the geographic locations of each village and non-village location. They then navigated to their starting location and scheduled interviews with four nearby households for a total of 40 households in each cell/cell cluster. Probably could have done more re: how the enumerators chose those households as well, but we left it up to them and their best judgement at that point. In the case that no head of household was home, the enumerators would attempt to find the head of household in the field; if no one could be found, or no appointment could be made, the enumerator would mark the location and move on to the next closest house.

During the survey the following day, enumerators marked the GPS location of each household they interviewed. At the close of each day, these locations were returned to me for subsequent use in analysis, along with the survey results for each household (which were uploaded directly into the online database).

Statistical Analysis

Our sampling approach described above results in a statistically dependent, unequal probability of selection sample. Inference based on the sample requires the use of sampling weights that reflect their non-equal and statistically dependent probabilities of inclusion. A general approach to this problem – developed for any sample generated without replacement – is the Horvitz-Thompson estimator [4, 2]. The total population estimator is,

$$\hat{\tau}_\pi = \sum_{i=1}^{v} \frac{y_i}{\pi_i}$$

where π_i is the probability of inclusion. The variance estimator for the total is,

$$\hat{var}(\hat{\tau_\pi}) = \sum_{i=1}^{v} (\frac{1}{\pi_i^2} - \frac{1}{\pi_i}) y_i^2 + 2 \sum_{i=1}^{v} \sum_{j>i} (\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}).$$

The mean and variance for a mean are derived from the totals. Importantly, the variance estimator requires not only own probabilities of inclusion but also joint probabilities of inclusion for every pair of observations included. In practice these joint probabilities can rarely be derived analytically and there are simplified estimators in the literature that impose assumptions [4, 2].

In our application, we estimate the own and joint probabilities of inclusion through repeated calls to our sampling algorithm. [explain resampling approach in detail]. Figures 1 and 2 visually depict the resulting inclusion probability estimates. Own probabilities are on the diagonal (lower left to upper right) and darker colors indicate higher probabilities of inclusion. Note the block diagonal structure indicating correlated probabilities for observations from the same cell. The partitions indicate observation blocks from formal villages (lower left) and informal villages (upper right). The upper left and lower right partition blocks contain cross-inclusion probabilities between formal village and informal village observations. Note again the slightly higher cross-inclusion probabilities in the diagonals. This is because of co-occurrence in the same cell or set of proximate cells.

Prior to the conference we will use a simulation study to show that our approach to estimating weights results in unbiased inference. The simulation study still needs to be completed.

Discussion

Survey sampling of households in rural Africa is difficult because national statistical agencies typically only have spatial data at aggregate zonal scales. If the object is to sample at the village level, only sampling along known (primary and secondary) roadways is likely to yield a sample biased towards larger villages. In our Rwandan study this was particularly problematic because we were trying to compare formal, government initiated villages (imigidugudu) to informal villages. The latter may only be accessible down a dirt path. Our sampling approach allows us to randomly select villages across the full landscape. The resulting unequal probability of selection sample requires the use of complex weighting for inference. We describe the sampling process and the construction of weights used for our inference.

References

[1] Rebecca Awuah et al. “An adaptive household sampling method for rural African communities”. In: *African Journal of Food, Agriculture, Nutrition and Development* 17.1 (2017), pp. 11477–11496.

[2] Tamie Henderson. "Estimating the Variance of the Horvitz-Thompson Estimator". PhD thesis. Australian National University, 2006.

[3] Michelle C Kondo et al. "A random spatial sampling method in a rural developing nation". In: *BMC public health* 14 (2014), pp. 1–8.

[4] Steven K Thompson. *Sampling*. Vol. 755. John Wiley & Sons, 2012.

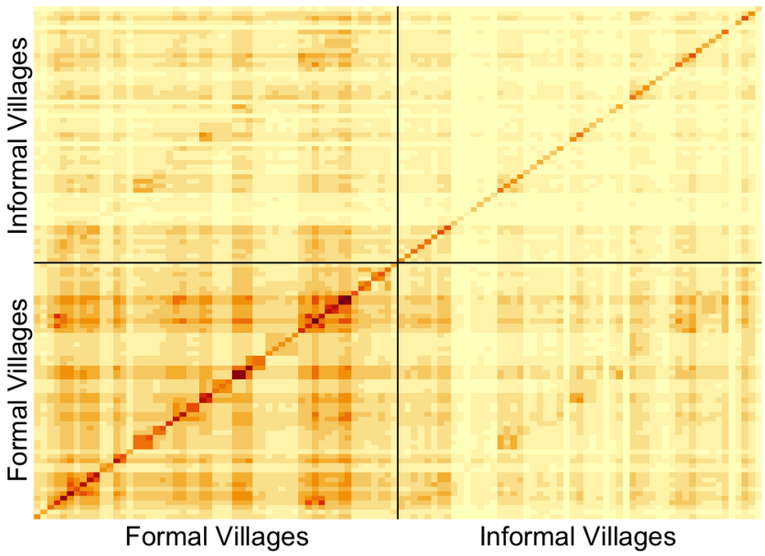

Figure 1: Nyamasheke: Own and cross inclusion probabilities based on resampling. Partitions indicate blocks of observations in formal villages and informal villages.

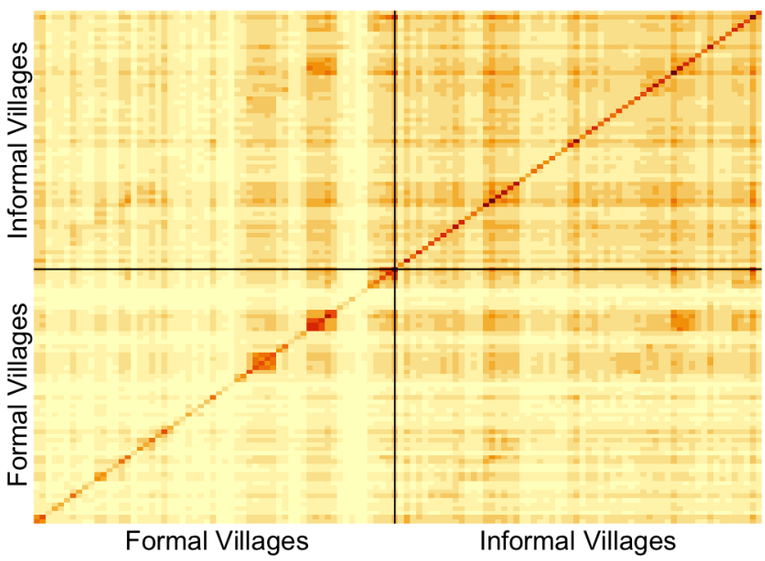

Figure 2: Musanze: Own and cross inclusion probabilities based on resampling. Partitions indicate blocks of observations in formal villages and informal villages.