7

Methods of Geo-Spatial Sampling

Stephanie Eckman and Kristen Himelein

1 Introduction

Technological advances in geospatial data have the potential to change how survey data are collected. Long hampered by high costs, limited capacity, and difficulties in supervision, sample selection is often done using second-best or nonprobability approaches. As geospatial technology has improved and become more widespread, costs have come down and the number of available tools have increased, making Geographic Information Systems (GIS)-based sampling approaches accessible to more users. This chapter presents experiences with GIS-based sampling from three different settings: (i) where no sampling frame is present because the census is outdated; (ii) sampling pastoralist communities; and (iii) rapid listing of enumeration areas to reduce exposure of field

S. Eckman
RTI International, Washington, DC, USA

K. Himelein (✉)
World Bank, Washington, DC, USA
e-mail: khimelein@worldbank.org

teams. The case studies focus on extreme situations, particularly those in conflict-prone areas, as innovation often takes place when few other options are available. The applications discussed here, however, are applicable to many less extreme situations.

2 Data Challenge and Innovation #1: Creating a Sampling Frame in the Absence of a Census

For many studies, no sampling frame of the target population is available. The most common approach to addressing this problem for large-scale household surveys in the developing world is to use a stratified two-stage design. In the first stage, census enumeration areas are selected as the Primary Sampling Unit (PSU), using probability proportional to estimated size. In the second stage, a household listing operation is conducted in the selected PSUs, and households are selected using simple random sampling.^{1,2} With this approach, even outdated census data can be used to select PSUs, as long as a high-quality listing operation is done in the selected PSUs to create a sampling frame for the second stage selection of households. Using out-of-date census data as a measure of size in PSU selection will result in estimates that are inefficient but still unbiased. However, some countries do not have census records at all because of accessibility issues, war, or natural disasters. In these situations, newly available high-resolution satellite data can be used to generate estimated population densities and to demarcate PSU boundaries. The two examples discussed here are from surveys conducted in rural Somalia and Kinshasa, Democratic Republic of the Congo (DRC).

¹For the purposes of this chapter, the word “dwelling” is used to denote a physical structure inhabited by one or more households, while a “household” is a group of individuals that function as an economic unit. All methods that select dwellings which have the possibility of containing multiple households have selection protocols to randomly select an individual household for interview.

²Grosh and Muñoz (1996).

In Somalia, the last population census, carried out in 1975, measured the population at 3.9 million. Current estimates for the country indicate a population of more than 14 million. For the DRC, similarly, the last census was carried out in 1984, at which time the population was around 29 million. Current population estimates are now over 77 million. As noted above, it would still be possible to use the outdated census for estimated population totals if there was an expectation of approximately constant growth across regions. Both Somalia and DRC, however, have experienced significant civil strife, including large-scale displacements of the population.

Some countries, notably Haiti following the 2010 earthquake, have used “quick counts” to collect information about where the population lives and to estimate its size. In a quick count, enumeration areas are randomly sampled and listed, then the results are used to build a model to update census counts in the remaining areas.³ However, in Haiti, the most recent census was only seven years old at the time of the earthquake, and the damage and population movements were relatively concentrated. The more time that has elapsed since the last census, the more difficult it is to develop an accurate model of the current population based on quick counts. Moreover, the DRC has a land area nearly 85 times the size of Haiti, which makes using a quick count methodology impractical from the perspective of both cost and implementation time. In Somalia, in addition to ongoing insecurity in certain areas, the enumeration area estimates from the 1975 census were never published, and the full results are thought to be lost. Therefore, alternative approaches to selecting a household sample were needed in both Somalia and the DRC.

2.1 Innovations

Three approaches were implemented across the two surveys. For the Somali High Frequency Survey (SHFS), rural areas posed a challenge for the creation of a sampling frame. Rural areas were defined as non-urban permanently settled areas but excluding Internally Displaced

³IHSI et al. (2012).

Persons (IDP) settlements.⁴ To create a frame for the first selection stage, a gridded population approach was developed in collaboration with Flowminder.⁵ Rural areas that were secure enough for data collection were divided into 100 by 100 meter grid cells. For each cell, WorldPop data provided an estimated population size.

Neighboring cells were then combined to form PSUs, using a quadtree algorithm, which combines cells to meet specified criteria, in this case, area and population size.⁶ The maximum area was set at 3 by 3 kilometers, and the maximum population was limited to 3500 to keep enumeration areas manageable for field teams. The left panel of Fig. 1 shows the PSUs created by the above steps, with the color indicating the estimated population in each one.⁷ Next, a sample of PSUs selected using probability proportional to estimated size. The selected PSUs were then further subdivided into segments. If the selected PSU contained 12 or fewer dwellings based on satellite imagery, only one segment was defined. For those PSUs containing between 13 and 150 dwellings, 12 segments were defined, with additional segments being defined for PSUs with more than 150 dwellings.

A major disadvantage of the grid approach described above is that the boundaries of the resulting PSUs do not follow natural boundaries such as roads, valleys, and rivers. The cells' artificial boundaries complicate field implementation. Aware of this constraint, the team initially pursued an alternative methodology in which the WorldPop distribution was used to randomly select points to serve as "seeds" for PSUs, which were then grown until they reached an estimated population of around 150 dwellings but without crossing natural boundaries.⁸ Unfortunately,

⁴In urban areas, boundaries and population estimates were available from the United Nations Population Fund's Population Estimation Survey. Boundaries of IDP settlements were provided by United Nations High Commission for Refugee's Shelter Cluster.

⁵Closely following the methodology by Muñoz and Langeraar (2013).

⁶See Samet (1984), for a description of the methodology, and Minasny et al. (2007), for an application of the methodology to sample design.

⁷The map shows both urban and rural areas. Urban areas were not subject to the same population or land area limits.

⁸Thomson et al. (2017).

*Population for construction PSUs in Somalia**Building classification in the city of Kindu, DRC*

Fig. 1 Building classifications (Color figure online) (Source Authors' calculation)

two major drawbacks became immediately apparent: the development of algorithms to detect natural boundaries was expensive and time-consuming, and selection probabilities were not straightforward to calculate because of boundary effects (seeds near boundaries could grow in fewer directions than others). The team therefore reverted to a gridded approach but manually adjusted segments to follow natural boundaries to mitigate potential implementation issues.

In the DRC, two methods were used. In the districts of Kisenso, Kimbanseke, and Mont Ngafula in Kinshasa, and the sites of Kindu, Tchonka, and Basankusu, a one-stage sample of dwellings was selected based on counts of dwellings made from satellite images. In partnership with the firm Satplan Alpha, the project used recent satellite images to count and geo-locate all dwelling units. This work was done manually. Team members classified each building in the satellite images as low-density residential, high-density residential, or non-residential, using their local knowledge of the typical characteristics of dwelling units in the DRC. These typical characteristics were locally specific,

varying between cities and between dense inner-city districts, peri-urban zones, and semirural areas on the outskirts. The main characteristics used to classify structures were architecture, building size and features, roof segmentation, roof design intricacy and height, building orientation, site boundary features, proximity to major streets, street activity, and traffic. The right panel in Fig. 1 shows the final map for Kindu, DRC, with each building classified as low-density residential (blue), high-density residential (yellow), and non-residential (red). When the counting, geo-locating, and classification were complete, each dwelling was assigned a random number, and a sample was selected through a one-stage random draw. If the classification was correct, this approach resulted in an equal-probability simple random sample of dwellings.

In the districts of N'djili and Makala in Kinshasa, a two-stage random sample was used.⁹ PSU boundaries were first defined using administrative and physical boundaries such as rivers, highways, and secondary and residential roads that would be easily identifiable by interviewers on the ground. The delineation process used an automated iterative approach where PSUs were created and then split or merged based on target population size. The left panel of Fig. 2 shows a map indicating the manually created PSUs.

The next step was to estimate the population within each of these PSUs from high-resolution satellite data. First, a Random Forest Regression model was used to estimate population density based on contextual image information (image metrics that incorporate various aspects of surrounding information, rather than single-pixel signature).¹⁰ The model was trained using a sub-sample of building locations.¹¹ The area and average building density for each PSU was then integrated with land use and land-cover data to adjust the area by the percentage covered with vegetation and then to produce a building count.¹²

⁹For further information, see Hirn and Rodella (2017).

¹⁰Implemented using MapPy, a Python library for remote sensing developed by Jordan Graesser.

¹¹Graesser et al. (2012) provides a more detailed description of contextual image information in image processing.

¹²Building counts derived in this way produce comparable results to manual rooftop counts.

PSUs defined by natural and administrative boundaries

Algorithmic estimate of population density

Fig. 2 Boundaries and population densities (*Source* Authors' calculation)

PSUs were selected with probability proportional to this estimated size. A full listing operation was then conducted in the selected PSUs prior to the second stage selection of households. This approach leads to estimates with larger variances, and therefore less precise estimates, than the single-stage approach because the resulting sample is clustered.¹³

2.2 Key Results and Implementation Challenges

Each of the methods described above produced a sampling frame from which a representative sample was selected. There were, however, substantial challenges in Somalia. For the SHFS, 407 PSUs were selected for the survey (320 urban and 87 rural), and 366 PSUs were selected as replacements (251 urban and 115 rural). After selection, the PSUs were overlaid with satellite imagery from Google Earth and Bing to verify the presence of dwellings. Following that process, 53% of rural PSUs and 2% of urban PSUs were discarded and replaced due to having no visible

¹³Eckman, S., and B. West (2016), "Analysis of Data from Stratified and Clustered Surveys," in Wolf, C., Joye, D., Smith, T., and Fu, Y. (Eds.), *Handbook of Survey Methodology*. Thousand Oaks: Sage, 477–487.

population. In some cases, it was necessary to replace a PSU multiple times before one with visible dwellings was identified.

The approach used in the DRC generated more reliable results. Both the single-stage and multi-stage methods yielded results close to what the interview teams found during the listing exercise. The single-stage approach, which manually located dwellings based on satellite imagery and then drew a one-stage random sample, was applied in three large districts of Kinshasa. Locating individual dwellings on satellite imagery remains a manual task that is both relatively time-consuming and cannot be entirely standardized. While guidelines can be set for identifying dwellings, in practice, judgment calls are often required to (for example) distinguish businesses or separate conjoined structures into multiple dwellings. When selected structures turned out to be businesses, empty or destroyed houses, or other non-dwelling structures, the misidentified structures were replaced by a randomly selected replacement dwelling. If such misidentification is not excessive and does not systematically vary across the sampled area, the sample can be assumed to remain unbiased. However, misidentification can increase costs and needs to be monitored closely. Systematic variation in the misidentification of households across the sampled area may bias the sample (for example, underrepresenting areas with many high-rise buildings if the true number of dwellings within high-rises is systematically under-identified in a rooftop count). From a practical point of view, interviewers also sometimes struggled to find the selected households in dense areas, because no addresses were available, only a rooftop view with a GPS point. This drawback can be mitigated, however, by equipping interviewers with GPS-capable phones and clear walking maps that point out local landmarks and house characteristics to help with identification.

The second approach used in the DRC, which first defined PSUs and then algorithmically estimated population numbers to allow for an unbiased two-stage selection, posed different challenges. First, refining the algorithm that estimates population density is technically more complex than a simple visual count of dwellings based on satellite imagery. Once in place, however, it can quickly create automated population estimates for large areas. A second challenge is the loss of statistical efficiency inherent in the two-stage approach. Third, interviewers carrying out the listing within selected PSUs sometimes struggled to

follow PSU boundaries and to distinguish which buildings were within or outside a given PSU. To minimize such problems, it is critical to prepare clear walking maps for interviewers and guidelines on how to deal with overlapping properties.

In the 28 PSUs in the Makala municipality in the Funa district of Kinshasa, both manual counting of residential buildings (the first method) and the modeling approach (the second method) were used, permitting a comparison between the two methods and the actual number of households identified in the field listing. Compared to an actual total of 9322 households recorded by the listing, the manual approach identified 7489 dwellings, while the modeling approach generated 10,667 dwellings in the same area. The correlations between the estimated and the actual values at the PSU level were 88.7 and 93.1% for the manual approach and the modeled approach, respectively. This important result indicates that the algorithm outperformed manual counting, at least for this application. See Fig. 3 for

Fig. 3 Listing totals, modeled estimates, and rooftop counts for Makala (Source Authors' calculations)

a comparison of the dwelling counts estimated by the two methods with the household totals generated in the listing operation, for the 28 PSUs in Makala.

3 Data Challenge and Innovation #2: Sampling Pastoralist Communities

Livestock ownership serves a diverse set of functions in the developing world, from food source to savings and use as an investment vehicle. The pastoralist sector, however, has recently come under increasing pressures from several sources, including an increased demand for meat and dairy products from expanding middle classes, climate change, and the loss of traditional pasture land to development. Those who are among the most vulnerable to these pressures are nomadic and semi-nomadic pastoralist populations, but the transitory nature of their living situation also hampers the collection of high-quality representative data on which to base analyses.

Because many pastoralists lack a permanent dwelling, they are excluded by a traditional two-stage sampling approach. In July and August 2012, the World Bank undertook a survey in the Afar region of Ethiopia to test a novel approach to sampling the general population, including pastoralists.

The Afar region was selected for the pilot project for several reasons. First, the World Bank has an ongoing relationship with the Ethiopia Central Statistics Agency (CSA), including supporting the implementation of the Ethiopia Rural Socioeconomic Survey, which includes a module on pastoralist issues. The CSA also has a high-quality existing GIS infrastructure and a relatively high level of training compared to other potential study areas. Third, the Afar region offered geographic advantages over other pastoralist areas. It covers a land area of approximately 72,000 square kilometers in the north-east of Ethiopia and is relatively isolated. Well-guarded national boundaries, geographic features, and traditional ethnic hostilities limit the migration of the Afar people outside the boundaries of the region.

3.1 Innovations

The approach used to ensure pastoralist populations were included was the Random Geographic Cluster Sampling (RGCS) method. In an RGCS design, points (latitude and longitude) are randomly selected, and then a circular cluster of a given radius is created around the central point. All eligible respondents found within this cluster are selected for the survey. The main advantage of this design is that it captures everyone who is inside the selected circle at the time of the survey, including those who do not have a permanent dwelling or who are temporarily away from their usual dwelling. Properly implemented, this design eliminates the underrepresentation of mobile populations. Similar methods are commonly used in both developed and developing world contexts to measure agricultural production and livestock.¹⁴

To increase efficiency and lower fieldwork costs, the Afar region was divided into five strata, defined by the expected likelihood of finding herders and livestock. Spatial datasets describing land cover, land use, and other geographical features were used as input to delineate five discrete, mutually exclusive strata. The first stratum consisted of land in or near towns; the second stratum consisted of land under permanent agriculture; the third stratum was considered to be the most likely to contain livestock, and consisted of land within two kilometers of a major water source, including the Awash River and its permanent tributaries, and which also met the criteria for pasture based on a vegetation index; the fourth stratum consisted of land between two and ten kilometers from a major water source which also met criteria of pasture land; and the fifth stratum consisted of the remainder of the land area, which was considered to have the lowest probability of finding livestock (Fig. 4).

A total of 125 points were selected from these five strata for the survey. The number of selected points was higher in the strata with the highest expected concentrations of potentially nomadic households and livestock (Stratum 3) and lower in areas of lower expected

¹⁴For a more complete list of previous applications, see Himelein et al. (2014).

Fig. 4 Stratification map

density (Stratum 5). The radii for the circles also varied across the strata. In areas with higher expected densities, smaller circles were used to keep the workload manageable. In areas where few or no livestock were expected, the circle radius was expanded to the largest feasible dimensions to maximize the probability of finding animals. Table 1 lists the definition, sample size, and radius used in each of the five strata.

Table 1 Stratification of the Afar region

Stratum	Description & expected likelihood of finding individual/livestock	Radius (km)	Points selected	Total area (km ²)	Percentage of total landscape
1	High likelihood: towns	0.1	10	33	<1
2	Almost no possibility: settled agricultural areas/commercial farms	0.5	15	930	2
3	High likelihood: within 2 km of major river or swamps	1	60	3538	6
4	Medium likelihood: within 10 km of major river or swamps	2	30	6921	12
5	Low likelihood: all land not in another stratum	5	10	45,152	80
Total			125	56,574	100

After the selection of the PSUs, teams were given maps and hand-held GPS devices to conduct the surveys. Upon arriving at the center of a circle, the team canvased the circle and interviewed all households within its boundaries. The GPS device showed the selected circle, and alerted interviewers when they crossed into or out of area.

3.2 Key Results

The pilot project of the RGCS technique to collect livestock data in the Afar region of Ethiopia demonstrated that the implementation of such a design is feasible. Of the 125 points selected, 102 were visited. Of those visited, 59 circles (58%) contained at least one livestock animal. In total, the interviewers collected information from 793 households that owned livestock, although nine of these households were shown by their GPS coordinates to be outside of the circle boundaries and were therefore excluded from the analysis, leaving a total sample size of 784. The number of interviewed livestock-owning households per circle ranged from one to 65, with a mean of approximately 15. In total, 3698 individuals living in households owning livestock were identified as part

of the survey. Of these, 127 reported having no permanent dwelling, which is a weighted estimate of 4701, or 2% of the livestock-holding population in the study area. All but five of the individuals without a permanent dwelling lived in households in which all members were completely nomadic. The inclusion of households without permanent addresses in the survey was a primary objective of the original research agenda because this group is traditionally underrepresented in dwelling-based surveys.

Overall, the project showed that sufficient GIS information is available, often in the public domain, to create strata for the probability of finding livestock, and to select points within those strata. With maps and relatively inexpensive GPS devices, interviewing teams can navigate the selected circles and identify eligible respondents within these clusters. The identified respondents can then be interviewed regarding their household's socioeconomic conditions and livestock holdings, creating the linkages necessary to understand the socioeconomic situation of these populations. In addition, using standard statistical methods, it is possible, although challenging, to calculate weights that take into account the varying probabilities of selection and that sufficiently address overlap probabilities. Moreover, information generated as part of the GPS field implementation can be used to account for underrepresentation, as discussed below. Finally, the methodology did what it was designed to do: Capture households without permanent dwellings that would have been excluded from a traditional dwelling-based sample design. The identification and interviewing of these households proved to be a major benefit to the RGCS, compared to the traditional household-based approach to survey sampling.

3.3 Implementation Challenges

Because the study area encompasses some of the harshest terrains in the region, and the methodology was novel for both the research and implementation teams, several unexpected difficulties were encountered. First, seasonal rains started earlier than expected, which created access problems such as the flooding of roads and land bordering the rivers.

The access issues necessitated longer walks for interviewers, including one incident where a team had to walk 15 kilometers to reach the selected site. Other physical obstacles such as national park boundaries, active volcanoes, and militarized areas further restricted access to some locations. Third, ongoing strained relations between local communities and the national government led to a few isolated security incidents, including minor assaults against drivers and fieldworkers, and the (brief) kidnapping of the survey coordinator.

Beyond the implementation challenges, two other substantial issues arose as part of the data analysis process. The first was related to the calculation of the weights, which was much more complicated than originally anticipated.¹⁵ The second challenge related to interviewers not canvassing the entire circle, and therefore missing potentially eligible respondents. The Viewshed analysis in Fig. 5 shows the path covered by the interviewers (the white lines), the portions of the circle they could have observed during their work (green and brown terrain map), and the black squares are the areas the interviewers could not have observed based on their path of travel. Several explanations for interviewers' failure to cover the entirety of the assigned circles are possible. The weather was extremely hot during this period. Flooding made access more difficult by requiring interviewers to take long detours on foot or ford swollen rivers. The survey took place during Ramadan, which limited the availability of local guides to assist the teams. Alternatively, however, it is feasible that the areas not observed were missed because they could not possibly contain any livestock, for example, because of the presence of flood water or vegetation too thick to traverse. Thus, the areas might be missing at random or not at random, and these two possibilities require different treatment in the analysis. Because it was impossible to distinguish the cause for the missed areas, two sets of statistics were reported for this study. This issue should be investigated closely for future implementations using this method.

¹⁵A full discussion of the correct procedure to derive probability weights is included in Himelein et al. (2014).

Fig. 5 Viewshed analysis (Color figure online)

4 Data Challenge and Innovation #3: Rapid Listing of Enumeration Areas¹⁶

The main challenge encountered in the Mogadishu High Frequency Survey (MHFS) Pilot¹⁷ was related to security issues, which made traditional listings of households within PSUs impossible. The MHFS was conducted between October and December 2014 by the World Bank and Altai Consulting. In this case, the PSUs were selected from existing census enumeration area maps using probability proportional to estimated size according to the United Nations Population Fund's Population Estimation Survey. In the second stage of the survey, however, carrying out a full listing was deemed unsafe. Listing households in a PSU would require the team to spend an entire day in one

¹⁶See Himelein et al. (2017) for more complete discussion of the context and analytical approaches, as well as for the complete set of results.

¹⁷The MHFSA is a different survey than the Somalia High Frequency Survey discussed in Sect. 1.1.

neighborhood, moving in a predictable pattern to reach all dwellings. The team's prolonged presence on the ground would increase their exposure to robbery, kidnapping, and assault, and increase the likelihood that local militias would object to their presence. A random walk procedure was initially proposed as a replacement, but this method has been shown in the literature to have a high likelihood of generating biased results, even if implemented under perfect conditions.¹⁸

The team considered four alternatives to a random walk. The first was to use satellite mapping to count rooftops. This methodology is shown in the right panel of Fig. 1 and discussed as the one-stage method used in the DRC survey above. The second alternative was segmentation, also shown in the left panel of Fig. 2: the creation of clusters with discernible boundaries on the ground. The third, grids, is discussed above. The fourth alternative was a novel proposal based on a random point selection methodology, but one that considers differing probabilities of selection generated by the spatial distribution of dwellings within a PSU.

Because Mogadishu at the time was deemed too dangerous to conduct pilots of the different methodologies, a comparison between the methods was made using a simulation study. The study simulated repeated sampling via the five methods described above in three purposefully chosen PSUs which varied in size, population, and socioeconomic status. Figure 6 illustrates the size and location of the selected PSUs.

To simulate the sensitivity of each method to different degrees of clustering, three methods were used to assign consumption values to the dwellings. In the first approach, values were randomly drawn from the distribution and assigned to dwellings in each PSU, resulting in no clustering in the consumption values. In the second and third approaches, the same values were reassigned within each PSU to create a moderate and a high degree of spatial clustering. After assignment, each dwelling in each of the three PSUs had three assigned consumption values.

¹⁸Bauer (2014, 2016).

Fig. 6 Size and location of selected PSUs

4.1 Innovations

Several surveys have used random point selection methodologies to select households. In these methods, a random starting point is selected, and the interviewer is instructed either to interview the nearest dwelling or to proceed in a set direction until a dwelling is reached. The main drawback of these approaches is that the weights are difficult to calculate. Many researchers assume that the resulting sample is equal probability,¹⁹ but that is not the case. A dwelling in a large open space has a higher probability of selection than one located in a densely-populated area: More points lead to the selection of the isolated dwelling.

The innovation proposed as part of the Mogadishu survey was to calculate the size of the “shadow” of the dwelling and use this information

¹⁹For further discussion, see Grais et al. (2007) and Kondo et al. (2014).

to estimate the probability of selection. Interviewers were instructed to travel to each preselected point within the PSU, walk in the direction of the Qibla (the direction of Mecca), and to interview the first dwelling they reached. They repeated this approach until a sample of 10 dwellings had been achieved. The Qibla was used in Mogadishu because many interviewers have an app on their cell phones that indicates this direction, but any verifiable direction (north, south, etc.) would work similarly well. The probability of selection of each dwelling is proportional to the size of its “shadow”: the set of all possible points that would lead to the selection of that dwelling. Figure 7 provides a visual representation of a dwelling’s shadow in the Qibla method. Other random point selection methods lead to differently shaped shadows, but the principle is the same.

A major potential drawback of the Qibla and other related methods is the difficulty of measuring the area of the shadows. If high-quality, up-to-date satellite maps exist, then it is possible to use these images to calculate the shadow of a dwelling. The size, however, would be distorted if new structures had been built or demolished since the image was taken. Calculating the area of the shadow in the field could possibly be done by

Fig. 7 Example of the Qibla method

asking the interviewer to walk the perimeter of the shadow with the GPS, but this would require substantial training, and may lead to measurement error. It would also increase the time spent in the field, which was not an option in an insecure context like Mogadishu. Two alternatives were therefore used to develop a proxy for the size of the shadow: The distance to the next structure in the opposite direction to the Qibla multiplied by the actual width of the dwelling, and the measured distance to the next structure multiplied by a categorical shadow width variable (small/medium/large) as defined by the interviewer. In addition, the simulation tested an approach which ignored the probabilities of selection and assumed the Qibla method led to an equal probability sample of dwellings.

4.2 Key Results

Figure 8 presents the results from the simulations of the five sampling methods. In this figure, the three PSUs are combined, but the sampling methods are shown separately. The points are the means of the sampling distributions

Fig. 8 Mean and confidence intervals (by method)

and the bars indicate the 5th and 95th percentiles. For each sampling method, there are three results shown: one for random assignment of consumption to dwellings (that is, no clustering); one for some clustering; and one for high clustering. The horizontal line at 40 is the true population mean consumption level. The results allow us to compare the methods' performance in terms of bias and variance and robustness to clustering.

The satellite method is unbiased: The mean over all the samples is the same as the true mean. This method is also unaffected by clustering in the consumption variable. These results were expected, because this method was assumed to be equivalent to the gold standard method of a full in-field listing; that is, the images were assumed to be up-to-date. Segmentation also showed consistently unbiased results, but higher variances for higher degrees of clustering in the underlying distribution, which is consistent with sampling theory on clustering.²⁰ The grid method, despite being conceptually similar to segmentation, overestimated the means with a bias up to 10% for the clustered distributions. The bias is related to grid squares that did not have enough dwellings to meet the sample size.²¹ As expected, the Qibla method with the correct weights yielded unbiased results, but with wide confidence intervals, although these were partially driven by a few outliers. The values of the 5th and 95th percentiles of the distribution for this method are similar to those in the segmentation method when clustering is applied. The two methods of estimating the measure of size for the Qibla method showed a small amount of bias, ranging between 1.5% and 6.5%, depending on the degree of clustering. The final Qibla alternative, the unweighted version, consistently underestimated the true mean. The random walk approach, as noted above, is not theoretically unbiased, and this is reflected in the simulation results.

The main lesson learned from the simulation experiments is that full listing and satellite mapping generate the most consistently precise and unbiased results. It is also possible to generate unbiased results using a random point selection method—in this case, the Qibla method—but this approach

²⁰See Eckman and West (2016).

²¹For a full discussion, see Himelein et al. (2017).

requires the accurate calculation of the area of the shadow to generate correct probability weights. If such complete data were available to researchers, satellite mapping would likely be a better choice. The Qibla method and segmentation are both unbiased and offer roughly similar precision, and therefore any choice between them would be based mainly on ease of implementation and the amount of information available. The other methods considered, including the proxy weights for the Qibla method and gridding, introduce some bias, but may be acceptable if other alternatives are not feasible. The two unweighted methods, the unweighted Qibla method and the random walk method, demonstrate the most bias, and should therefore be avoided.

5 Implementation Challenges

Because this study involved simulations and no fieldwork, fewer challenges were encountered. As discussed above, the main implementation challenges encountered for the Qibla method were related to the calculation of the shadow area, and by extension, the sample weights. In addition, some issues were encountered when the pre-selected point did not lead to the selection of a dwelling within the boundaries of the PSU. This issue was more pronounced in PSUs that had more open space, particularly on the perimeter of the city. In an actual survey environment, fieldwork protocols and training would be necessary to ensure consistency in addressing these situations.

6 Lessons Learned and Next Steps

It is clear from the accelerating pace of the application of GIS-based technology to sample design that the field will continue to expand in the coming years, driven by less expensive and higher resolution imagery and the development of better algorithms. Despite the excitement that these advances generate, however, researchers and practitioners must not lose sight of the importance of calculating accurate probabilities of selection to generate unbiased estimates. As shown by the RGCS design and the Qibla method, these calculations can be

challenging, and there is a need for two complementary research areas in GIS-based sampling. The first area where research is needed is in improved population estimates where there is no census (or equivalent) frame. The work described above in the DRC and rural Somalia was a step in this direction. Flowminder, Facebook, WorldPop, and other groups have released population estimates.²² The second research area is in relation to new methods of household selection when listing is not possible. The experiments in Afar and Mogadishu offer two alternatives. Unfortunately, both led to potentially complex weight calculations and overly variable weights, which introduce variance into estimates. New technologies, such as unmanned aerial vehicles, also have the potential to reduce the time and costs involved in listing operations.²³

Cost is also an important consideration when deciding between traditional and innovative methods. Any non-traditional method will incur additional costs associated with preparation and training, but these will decrease over time as familiarity grows. For example, the DRC two-stage mapping exercise required the purchase of imagery costing \$10,000, as well as three weeks of work from an experienced GIS specialist (who was new to this specific image processing application; a specialist with experience in the mapping application could have done the processing in less time). Imagery could also be obtained less expensively, by, for example, using lower resolution images or free OpenStreetMap data, where available.²⁴ The costs of the new techniques must be weighed against the costs of listing, which increases data collection costs by approximately 25% in each cluster. However, the cost of using either type of methodology is lower than employing a non-probability design, which does not guarantee reliable or representative estimates, regardless of the cost of data collection.

Acknowledgements The authors gratefully acknowledge the comments and contributions of Maximilian Hirn, Siobhan Murray, Utz Pape, and Aude-Sophie Rodella of the World Bank, and Sarchil Qader of Flowminder.

²²For a further discussion, see Facebook Code (2017) and LandScan (2017).

²³See Eckman et al. (2018).

²⁴For further detail and availability, see openstreetmap.org.

References

- Bauer, J. J. (2014). "Selection Errors of Random Route Samples." *Sociological Methods & Research* 43 (3): 519–544.
- Bauer, J. J. (2016). "Biases in Random Route Surveys." *Journal of Survey Statistics and Methodology* 4: 263–287.
- Eckman, S., and B. West. (2016). "Analysis of Data from Stratified and Clustered Surveys." In Wolf, C., Joye, D., Smith, T., and Fu, Y. (Eds.), *Handbook of Survey Methodology*. Thousand Oaks: Sage, 477–487.
- Eckman, S., J. Eyerman, and D. Temple. (2018). *Unmanned Aircraft Systems Can Improve Survey Data Collection*. Research Triangle Park, NC: RTI Press. <http://goo.gl/eRHKpy>.
- Facebook Code. (2017). *Open Population Datasets and Open Challenges* [Online]. Available at <https://code.facebook.com/posts/596471193873876/open-population-datasets-and-open-challenges/>. Accessed 6 November 2017.
- Graesser, J., A. Cheriadat, R. R. Vatsavai, V. Chandola, J. Long, and E. Bright. (2012). "Image Based Characterization of Formal and Informal Neighbourhoods in an Urban Landscape." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5 (4): 1164–1176.
- Grais, R. F., A. M. Rose, and J. P. Guthmann. (2007). "Don't Spin the Pen: Two Alternative Methods for Second-Stage Sampling in Urban Cluster Surveys." *Emerging Themes in Epidemiology* 4 (1): 8.
- Grosh M., and J. Muñoz. (1996). *A Manual for Planning and Implementing the Living Standards Measurement Study Survey*. Washington, DC: World Bank.
- Himelein, K., S. Eckman, and S. Murray. (2014). "Sampling Nomads: A New Technique for Remote, Hard-to-reach, and Mobile Populations." *Journal of Official Statistics* 30 (2): 191–213.
- Himelein, K., S. Eckman, S. Murray, and J. Bauer. (2017). "Alternatives to Full Listing for Second Stage Sampling: Methods and Implications." *Statistical Journal of the IAOS* 33 (3): 701–718.
- Hirn, M., and A. S. Rodella. (2017). *WASH Poor in a Water-Rich Country: A Diagnostic of Water, Sanitation, Hygiene, and Poverty in the Democratic Republic of Congo*. WASH Poverty Diagnostic Series. Washington DC: World Bank. Available at <https://openknowledge.worldbank.org/handle/10986/27320>.

- Institut Haïtien de Statistique et d'Informatique (IHSI), Développement Institutions et Mondialisation, and World Bank. (2012). *Enquete sur les Conditions de Vie des Ménages Après Séisme (ECVMAS) 2012*. Available at http://www.ihsi.ht/pdf/ecvmass/ecvmass_metadonnees/0_ECHANTILLON/0_ECVMAS_Plan%20Echantillonnage_28052013.pdf. Accessed 1 November 2017.
- Kondo, M. C., K. D. Bream, F. K. Barg, and C. C. Branas. (2014). "A Random Spatial Sampling Method in a Rural Developing Nation." *BMC Public Health* 14 (1): 338.
- LandScan. (2017). *LandScan—Data Availability* [Online]. Available at http://web.ornl.gov/sci/landscan/landscan_data_avail.shtml. Accessed 6 November 2017.
- Minasny, B., A. B. McBratney, and D. J. Walvoort. (2007). "The Variance Quadtree Algorithm: Use for Spatial Sampling Design." *Computers & Geosciences* 33 (3): 383–392.
- Muñoz, J., and W. Langeraar. (2013). *A Census-Independent Sampling Strategy for a Household Survey in Myanmar*. Available at <http://winegis.com/images/census-independent-GIS-based-sampling-strategy-for-household-surveys-plan-of-action%20removed.pdf>. Accessed 2 November 2017.
- OpenStreetMap. (2017). OpenStreetMap [Online]. Available at <https://www.openstreetmap.org>. Accessed 20 November 2017.
- Samet, H. (1984). "The Quadtree and Related Hierarchical Data Structures." *ACM Computing Surveys (CSUR)* 16 (2): 187–260.
- Thomson, D. R., F. R. Stevens, N. W. Ruktanonchai, A. J. Tatem, and M. C. Castro. (2017). "GridSample: An R Package to Generate Household Survey Primary Sampling Units (PSUs) from Gridded Population Data." *International Journal of Health Geographics* 16 (1): 25.

The opinions expressed in this chapter are those of the author(s) and do not necessarily reflect the views of the International Bank for Reconstruction and Development/The World Bank, its Board of Directors, or the countries they represent.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 3.0 IGO license (<https://creativecommons.org/licenses/by/3.0/igo/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the International Bank for Reconstruction and Development/The World Bank, provide a link to the Creative Commons license and indicate if changes were made.

Any dispute related to the use of the works of the International Bank for Reconstruction and Development/The World Bank that cannot be settled amicably shall be submitted to arbitration pursuant to the UNCITRAL rules. The use of the International Bank for Reconstruction and Development/The World Bank's name for any purpose other than for attribution, and the use of the International Bank for Reconstruction and Development/The World Bank's logo, shall be subject to a separate written license agreement between the International Bank for Reconstruction and Development/The World Bank and the user and is not authorized as part of this CC-IGO license. Note that the link provided above includes additional terms and conditions of the license.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.