Second-Stage Sampling for Conflict Areas

Methods and Implications

Kristen Himelein

Stephanie Eckman

Siobhan Murray

Johannes Bauer

WORLD BANK GROUP

Poverty and Equity Global Practice Group

March 2016

Abstract

The collection of survey data from war zones or other unstable security situations is vulnerable to error because conflict often limits the implementation options. Although there are elevated risks throughout the process, this paper focuses specifically on challenges to frame construction and sample selection. The paper uses simulations based on data from the Mogadishu High Frequency Survey Pilot to examine the implications of the choice of second-stage selection methodology on bias and variance. Among the other findings, the simulations show the bias introduced by a random

walk design leads to the underestimation of the poverty headcount by more than 10 percent. The paper also discusses the experience of the authors in the time required and technical complexity of the associated back-office preparation work and weight calculations for each method. Finally, as the simulations assume perfect implementation of the design, the paper also discusses practicality, including the ease of implementation and options for remote verification, and outlines areas for future research and pilot testing.

This paper is a product of the Poverty and Equity Global Practice Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at <http://econ.worldbank.org>. The authors may be contacted at khimelein@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Second-Stage Sampling for Conflict Areas: Methods and Implications

Kristen Himelein, Stephanie Eckman, Siobhan Murray, and Johannes Bauer¹

JEL classification: C15, C83, F51, I32

¹ Kristen Himelein is a senior economist / statistician in the Poverty Global Practice at the World Bank. Stephanie Eckman is a senior survey research methodologist at RTI International in Washington, DC. Siobhan Murray is a technical specialist in the Development Economics Data Group in the World Bank. Johannes Bauer is research fellow at the Institute for Sociology, Ludwigs-Maximilians University Munich and at the Institute for Employment Research (IAB). All views are those of the authors and do not reflect the views of their employers including the World Bank or its member countries. The authors would like to thank Utz Pape and David McKenzie, from the World Bank, and Matthieu Dillais from Altai Consulting, for their comments on earlier drafts of this paper, and Hannah Mautner and Ruben Bach of IAB for their research assistance.

Introduction

The collection of survey data from war zones or other unstable security situations provides important insights into the socioeconomic implications of conflict. Data collected during these periods, however, are vulnerable to error, because conflict often limits the options for survey implementation. For example, the traditional two-stage sample design for face-to-face surveys in most developing countries first selects census enumeration areas as the primary sampling unit (PSU) with probability proportional to size and then conducts a listing operation to create a frame of households from which a sample is selected. Such an approach, however, may not be feasible in conflict areas. At the first stage, updated counts are often not available, making probability proportional to size selection inefficient. Also as the second stage requires that survey staff canvas the entire selected area, it may also be too dangerous in a conflict setting. As a result, many surveys of conflict areas are limited to qualitative work or resort to non-probability designs.

This paper uses simulations to explore several alternative sampling approaches considered for the baseline of the Mogadishu High Frequency Survey Pilot (MHFS). The baseline was a face-to-face household survey in Mogadishu, Somalia, conducted from October to December 2014 by World Bank and Altai Consulting. A full listing (see Harter et al, 2010 for details) was deemed unsafe in Mogadishu because the additional time in the field and the predictable movements by interviewers would increase their exposure to robbery, kidnapping, and assault, and increase the likelihood that the local militias would object to their presence. The survey needed an alternative second-stage sample design that would minimize the time spent in the field outside the households, but also could be implemented without expensive equipment or extensive technical training. In addition, international supervisors from the consulting firm could not go to the field, necessitating a sample design in which quality could be verified ex post.

The implementing partner originally proposed a random walk procedure. While this methodology has the benefits of fast implementation and unpredictability of movement, the method is non-probability and literature has shown the procedure to give biased results, even if implemented under perfect conditions (Bauer, 2014). Intuitively, a random walk would only be unbiased if the paths taken during the selection crossed each household once and only once, which is extremely unlikely in the field. Therefore the team considered four alternatives for household selection. The first option was to use a satellite map (of which many high quality options exist, due the limited cloud cover and political importance of the region) to identify all structures in the PSU and select ten for the survey. The second option considered was to subdivide the selected PSUs into segments consisting of eight to ten households and ask enumerators to list and choose households from the segments.² The segments would be of roughly equal size in terms of number of households but are likely to have irregular outlines reflecting the irregular layout of structures in Mogadishu. The third option considered was to lay a uniform grid over the PSU and ask enumerators to list and choose households from selected grid boxes. The final option considered was to start at a random point in the cluster and walk in a set direction, in this case the Qibla, or direction in which Muslims pray, until the interviewer encountered a structure.

² Interviewers had a selection application on their smart phones that they used whenever subsampling was needed.

The paper will make use of data from the completed Mogadishu pilot survey and geo-referenced maps and three example EAs to explore the following questions: (1) How might a given method be implemented in the field, given the information available and the security constraints? (2) What information is necessary to generate sampling weights, necessary for representative estimates, and how should those weights be calculated? (3) What are the implications in terms of precision and bias for each of the methods described above? (4) What are the implementation concerns for each method, including the options for verification and the impact of non-household structures?

The next section briefly describes the literature as it relates to the questions above, followed by section 2 describing the data. Section 3 addresses research questions 1 and 2 by giving further detail on the methods considered. Section 4 presents simulation results covering questions 3 and 4. Section 5 concludes by offering some discussion of the overall performance and potential future applications.

1. Literature Review

The most common method for collecting household data in Sub-Saharan Africa is to use a stratified two-stage sample, with census enumeration areas selected proportional to size in the first stage and a set number of households selected with simple random sampling in the second stage (Grosh and Munoz, 1996). Since administrative records are often incomplete and most structures do not have postal addresses, as is the case in Mogadishu, a household listing operation is usually necessary prior to the second-stage selection. However, due to the security concerns cited above, listing was not feasible.

A number of alternatives for second stage selection can be used when household lists are not available. A common alternative used in both Europe (see Bauer, 2014, for recent examples) and in the developing world is a random-walk. The Afrobarometer survey, which has been conducted in multiple rounds in 35 African countries since 1999, and the Gallup World Poll, which conducted surveys in 29 Sub-Saharan African countries in 2012, use random walk methodologies. Although the random walk methods do not necessarily produce equal probability samples, they do not collect any information with which to calculate probabilities of selection. For this reason, weights are not calculable for random walk samples; instead, the samples are analyzed as if they were equal probability. Bauer (2014) shows that this assumption is not correct by simulating all possible random routes using standard procedures within a German city and finds substantial deviation from equal probability. These results apply even when interviewers perfectly implement the routing instructions, which is unlikely given the limited ability to conduct in-field supervision of random walk selection and strong (though understandable) incentives for interviewers to select respondents who are willing to participate (Alt et al 1991). Several other studies have also shown that data collected via random walk do not match the population on basic demographics such as age, sex, education, household size, and marital status (Bien 1997, Hoffmeyer-Zlotnick 2003, Blohm 2006, Eckman & Koch 2016).

In the context of Mogadishu, a household listing was too dangerous and costly, a random walk too biased, and no household or person register existed. Therefore the researchers explored several alternative methods using a combination of satellite maps and area-based sampling. As satellite technology has improved in quality and become more readily available, it has been increasingly used for research in the developing world. Barry and Rüther (2001) and Turkstra and Raithelhuber (2004) use satellite imagery to study informal urban settlements in South Africa and Kenya, respectively. Aminipouri et al (2009) use

samples from high resolution satellite imagery to estimate slum populations in Dar-es-Salaam, Tanzania. Afzal et al (2015) incorporated satellite data into poverty prediction modeling for Pakistan and Sri Lanka, and concluded that its inclusion can lead to substantial improvements in modeled estimates of poverty.

Specifically related to sampling, the literature is more limited and mainly found in the public health literature. Dreiling et al (2009) tested the use of satellite images for household selection in rural counties of South Dakota, and found, while less time consuming, the method did a poor job of identifying the inhabited dwellings. Grais et al (2007) used a random point selection methodology in their study of vaccination rates in urban Niger, and compared the results to a random walk. They do not find statistically significant differences between the methods, though the sample size was limited, but conclude that interviewers found the random point selection methods more straightforward to implement than the random walk. Lowther et al (2009) used satellite imagery to map more than 16,000 households in urban Zambia to select young children for a measles prevalence survey. They find the method easy to implement, but do not do a formal comparison with alternatives. Kondo et al (2014) use a point selection mechanism in the city of Sanitiago Atitlán similar to our proposed Qibla method, but assume the method to be equal probability. A similar method was used by Kumar (2007), in which satellite maps overlaid with remote sensing data were used to create stratification for an air pollution study in India, and then selected random points. Kolbe and Hutson (2006) use a similar method to select households in Port-au-Prince, Haiti. Their study incorporates the probability of selection for randomly selecting from nearby structures when the selected points do not fall on the roof of a dwelling, but these probabilities would be distorted if the household density were unevenly distributed or for households close to the boundary. Other public health studies, such as the World Health Organization Expanded Program on Immunization studies, use the “spin the pen” method to choose a starting household and then interview a tight cluster of households, though this method has been shown to be nonprobability (Bennett et al 1994, Grais et al 2007).

Outside the field of public health, Himelein et al (2014) used circles generated around random points to survey pastoralist populations in eastern Ethiopia, with the stratification developed from satellite imagery. A variation of this method was considered for the High Frequency Survey Pilot, but the methodology is likely unsuited to a dense metropolitan area, because it involves surveying all households within the selected circles. The uncertainty over the final cluster size was also an issue in Mogadishu as clusters with too few households increased costs, while clusters with too many households increased the time in the field and raised security concerns.

This paper brings together alternatives developed from this literature and applies them to a conflict environment. We take a rigorous approach using simulations and careful estimation of weights to compare the methods across a variety of potential field conditions. The results offer general guidelines for practitioners developing implementation plans for conflict settings.

2. Data

To explore the challenges of the random walk and the four proposed alternatives, we simulated the use of each method in three example PSUs from Mogadishu, Somalia. We purposefully chose three census enumeration areas as the PSUs for this exercise to illustrate the variation in physical layout present in

Mogadishu. Maps of the three example PSUs are shown in the appendix. The first is in Dharkinley district, a comparatively wealthy section of southwestern Mogadishu where the households are laid out in relatively uniform gridded streets. This PSU has 68 total structures, and a total area of 24,390 square meters according to the December 2013 Google Earth imagery, which was the most current at the time of the initial analysis (in January 2015). The second PSU is on the eastern edge of Heliwa district in the northeast of the city. This area is more irregular in layout with larger gaps between buildings, and has a total of 309 structures in an area of 42,615 square meters based on imagery from March 2014. The third selected was in the more central Hodon district. It is densely populated with very irregularly laid out structures, has 353 total structures, and a total area of 345,157 square meters. This is also based on imagery from March 2014.

We explore the impact of each sampling method on estimates of household consumption. To construct the data set for the simulations, we drew consumption totals from the data collected by the MHFS. The survey covered both households in neighborhoods and those in internally displaced persons camps, but for the purposes of this simulation, we use only the neighborhood sample as within the camps there is little variation in consumption, due to reliance on food aid; furthermore, there are no camps within our three PSUs. Data were collected from the selected households on a limited range of food and non-food items which we sum to calculate a consumption measure (see Mistiaen and Pape, forthcoming, for further details on these calculations). There were 624 cases outside the IDP camps with non-missing values on the two consumption measures. The distribution of consumption across these cases is shown in Figure 1. The values follow a log-normal distribution and the underlying normal distribution has mean 40.0 and standard deviation 27.5.³

To simulate the variety of situations that may be found in the field, we use three different mechanisms for assigning consumption values to households in the three example PSUs. In the first, values are randomly assigned across the households in each PSU. In the second, the same values are reassigned to households to create a moderate degree of spatial clustering. In the third assignment mechanism, the spatial clustering of consumption values is more extreme. We study the ability of each of the proposed methods to estimate consumption under these three conditions. While these distributions may not mimic actual conditions in these PSUs, they are illustrative of the different situations encountered in the field.

³ It should be noted that since the values used in the simulations are drawn from the data collected, the data used in the simulations are subject to any bias due to non-response present in the data collection. While this is an issue for

3. Sampling Methodologies

For each of the methods (satellite mapping, segmentation, grid squares, Qibla method, and random walk) and assignment mechanisms (random assignment, spatial clustering, and extreme spatial clustering), 10,000 simulated samples were drawn and relevant probability weights calculated. Each sample consisted of ten structures per PSU. For the cases in which the household sample was selected in two stages, two segments of five households were chosen within each PSU for segmentation, and two grid boxes of up to five households were randomly chosen in the grid square selection. This section provides further detail on the selection methods and describes the weight calculations necessary to achieve unbiased results.

3.1 *Satellite mapping*

A full mapping of the PSU entails using satellite maps to identify the outline of each structure (see appendix). In this case, we used maps publically available on Google Earth and maps of the EA boundaries provided by the Somali Directorate of Statistics. From these maps, the structures inside each PSU can be assigned numbers (either by hand or digitally) and selected easily in the office with simple random sampling. The coordinates of the selected households can be loaded onto GPS devices to assist interviewers' locating efforts.

This approach is the closest of the proposed study methods to the gold standard of a well-implemented full household listing. The main differences are that in a field listing, enumerators can exclude ineligible structures, such as uninhabited and commercial buildings, and include information not available from satellite maps, such as the identification of individual units within multi-household structures. In addition, whereas listing is always done just before the data collection, the satellite maps may be out of date, leading to under-coverage of newly constructed units and/or selection of units that no longer exist. As noted in the previous section, the maps used for this paper were about one year old at the time of the initial analysis. The historical imagery from Google Earth indicates these maps are generally updated at least once a year, but this may vary substantially depending on the specific location and year. Selection from a satellite mapping therefore requires an additional set of field protocols for addressing and documenting the above issues.

The calculation of the probability of selection, and by extension the survey weight, is straight-forward. The probability is simply $\frac{n}{N}$, where n is the number of structures selected and N is the total number of structures, plus any necessary adjustment for multi-household structures (for example, one unit from a three unit building) encountered in the field.

3.2 *Segmenting*

Segmenting is a standard field procedure of subdividing large PSUs into smaller units, approximately equal sized in terms of number of households, for listing and selection purposes. The individual segments are

estimating the true mean in the population, it should not affect the means when compared between methods. First, as this bias is associated with a selected household's decision to participate, it would impact all methods equally. Secondly, as non-response attenuates any differences, they would appear smaller in magnitude in the simulations than in the true population. To address this, we ran a large number of simulations to generate narrow confidence intervals on the results.

then selected with simple random sampling, listed by field enumerators, and households selected from these lists. Segmenting is less time consuming than a full mapping exercise in terms of office preparation, but still requires the manual demarcation of segment boundaries. When creating segments, best practice is to use clearly discernable landmarks to draw boundaries, but these can change over time or not be correctly identified by the interviewers. If the interviewer incorrectly identifies the segment, it may be necessary to exclude the resulting data as they cannot be properly weighted. Properly implemented this method is able to produce unbiased estimates, but is not as dangerous or costly as a full listing, as listing only the selected segments would involve substantially less time in the field.

Figure 2: Example of Grid Sampling Method

The calculation of the probabilities of selection is also straightforward: it is the

Source: Authors' diagram based on PSU boundaries and Google Earth images

product of the probability of selection of the segment and the probability of selection of the household within the segment. The additional clustering introduced by this method, however, could decrease the precision of estimates due to the design effects. The magnitude of the decrease in precision would depend on the number of segments selected, the number of households selected per segment, and the degree of homogeneity within PSUs for the study variables. The impact would largest if all ten households were selected from the same segment, and decrease as more segments were selected. At the other extreme, if one household were selected from ten segments, segmentation would produce more precise estimates than simple random sampling as the segmentation prevents a chance geographic concentration of selected households. As a balance between efficiency and practicality, two segments and five households per segment were selected for the simulations.

3.3 Grid Squares

To implement the grid method, a uniform grid of squares (or other uniform shape) is overlaid on the PSU map. Figure 1 shows an example using 50 x 50 meter squares for the Dharkinley PSU. The area of a grid square includes all of the area that lies both within the grid square and within the PSU boundaries. For example, in grid square 17 in figure 1, the majority of the structures inside the square would not be eligible for the survey, as they lay outside of the PSU boundaries. Only the structures which lie in the bottom left corner are both within the grid and PSU boundaries.

One or more squares can then be selected with simple random sampling from the set of all squares that overlap the selected PSU. Depending on the survey protocols, a structure may be defined as eligible if all or part of it lies within the grid space. The more common protocol, including the structure if the majority lies within the grid square, has the benefit of simplifying the weight calculations, but the risk of subjective decisions made by interviewers in the field about where the majority of the building lies, which could lead to some buildings having no chance of selection. Since the options for supervision and field re-verification were limited in this survey, it was decided to consider the structure as eligible if any portion of the structure lay within the grid boundaries, to ensure that all units had a positive probability of selection.

To select a sample of households within the selected squares, a common approach would be for interviews to be conducted with all eligible respondents within the grid square. This could lead, however, to issues with verification as well as decreasing control over the final total sample size. Therefore the protocol used in Mogadishu had interviewers list all households within the selected squares and use the application on their smart phones to select a fixed number of households for the survey.

This variation of the grid method has the advantage that it requires less preparation time compared to mapping or segmenting. There are considerable drawbacks, however, in the ease of implementation and additional work to accurately calculate the selection probabilities. Since the grid squares do not follow visible landmarks, the boundaries must therefore be programmed into the GPS and identified by the interviewers. As it is unlikely that they will be able to walk straight along the boundary, additional training may be required to correctly identify eligible structures.

This approach also still requires some listing work, which may have security implications depending on the size of the squares in the grid. The size can vary depending on the physical size of the PSU and the density of the population. Larger grid squares may be necessary in sparsely populated areas, but increase listing time and interviewer exposure. Smaller squares require less listing work, but also mean that more buildings will lie on the boundaries between squares. Those selected structures which lie on boundary lines require either an arbitrary and unverifiable decision by field staff as to whether the majority of the structure lies within the grid square, or additional time for field implementation, as discussed below.

Let s be the number of squares selected in PSU and S be the number of squares that are partially or completely contained within PSU . For households that are entirely contained within square j , the probability of selection, given that PSU was selected, is:

$$\pi_{i|PSU} = \frac{n_j}{N_j} * \frac{s}{S} \quad (1)$$

where n_j is the number of structures selected from square j and N_j is the total number of eligible structures in the square. $\frac{s}{S}$ is the probability of selection of the square when a simple random sample of size s is selected from the S squares in PSU .

If household i lies in both squares j and j' , the probability of selection is:

$$\pi_{i|PSU} = \left(\frac{n_j}{N_j} * \frac{s}{S} \right) + \left(\frac{n_{j'}}{N_{j'}} * \frac{s}{S} \right) - \left(\frac{n_j}{N_j} * \frac{n_{j'}}{N_{j'}} * \frac{s}{S} * \frac{s-1}{S-1} \right) \quad (2)$$

For a structure overlapping more than two grid squares, there would be additional terms in equation (2), up to the extreme case lying on a four way intersection. Interviewers would also have to spend significant time on additional listing, which greatly increases exposure in the field and provides disincentives to interviewers to report such households.

3.4 Qibla Method

This sampling approach involves selecting multiple random locations within each PSU and traveling from each selected point in a common fixed direction until a structure is found. If the first structure the interviewer encounters is a household, the interview is done with the household. In Somalia, the consulting firm suggested using the Qibla (the direction in which Muslims pray) since it is common for interviewers to have an app on their cell phones which indicates this direction. Figure 3 gives a stylized example of this method. Household 510 will be interviewed whenever any of the points in the shaded region are selected. This region includes the area of the dwelling itself (its roof) and all points in its “shadow” – that is, all land inside the PSU that lies in the direction opposite the Qibla, excluding points that lead to the selection of other buildings. See figure A4 in the appendix for an example at the PSU level.

Despite its seeming ease-of-use, this approach contains many challenges. For one, it is not clear how non-residential structures should be handled. The interviewer could walk around business and vacant housing units, continuing in the Qibla direction until she finds a residential unit. This approach would work in theory, but in addition to the difficulties in remote verification it would create, it would also complicate the calculation of probabilities of selection (discussed below). Therefore we do not suggest it. Instead, we suggest coding points that lead to non-household selections as out-of-scope, and selecting additional points to replace them.

Perhaps the biggest challenge with this method is the collection of the information needed to calculate probabilities of selection of the selected households. Figure 3 shows Household 510 and, in the shaded area, the set of all points that lead to the selection of this household. Each household, i , in the PSU has an associated selection region: call this region A_i . The probability of selection of household i (conditional on selection of PSU), if c points in the PSU are selected, is one minus the probability that all c selected points are not in A_i :

Figure 3: Example of Qibla Method

Source: Authors' diagram based on PSU boundaries and Google Earth images

$$\pi_{i|PSU} = 1 - \left(1 - \frac{\text{Area of } A_i}{\text{total area of PSU}}\right)^c \quad (3)$$

(based on Särndal et al. 1992, p.50). This approach is essentially probability proportional to size selection with replacement, where the measure of size is the area of A_i . The weight is then the inverse, π_i^{-1} .

From Equation 3, the most difficult quantity to calculate is the area of A_i . For the purposes of this paper, we manually delineated building footprints individual structures from relatively recent Google Earth maps to calculate the A_i region for each selected household. Though requiring no additional software expertise, this method was time consuming in terms of preparation. For the three PSUs used in the paper, it took about one minute per household to construct a digital outline. If the PSUs contain approximately 250 structures (the ones used here contain 68, 309, and 353 structures, respectively), mapping the 106 PSUs selected for the full Mogadishu High Frequency Survey Pilot would have required more than 50 work days. It may be possible to automate the process for larger mapping efforts by using GIS-based algorithms for feature extraction that were not used here due to the limited number of PSUs.

Calculating A_i would be much harder, if not impossible, if high quality and recent satellite maps are not available. Any structures added since the imagery was captured would not appear and resulting areas of neighboring structures would be incorrectly included in A_i . We therefore also consider three methods of approximating A_i as defined in Equation 3. The first is the distance to the next structure in the opposite direction of the Qibla multiplied by the actual width of the dwelling (proxy weights 1). This is $l \times w$ in figure 4. The second is the measured distance to the next structure multiplied a categorical shadow width variable as defined by the interviewer (proxy weights 2), and the third ignores the weights completely. Though theoretically biased, the variations have the benefit of not requiring digitized maps and being more flexible in accounting for new construction, and under certain conditions they may be a good alternative for researchers who find themselves in second or third best scenarios.

The first alternative requires additional information from the field teams. Way points (latitude and longitude coordinates) must be captured with the GPS at the selected point and when the interviewer

arrives at the structure so that the distance can be calculated. Then the actual width of the structure perpendicular to the Qibla must be measured, which may be complicated if the dwelling has an irregular outline or if the perpendicular runs diagonally through the building. This may be done by asking interviewers to record a track as they walk the perimeter of the structure, though this requires additional processing from the team following data collection. The second variation is simpler to implement in that it does not

Figure 4: Calculation of Proxy-Weight Qibla Method

Source: Authors' diagram based on PSU boundaries and Google Earth images

involve any additional measurements, beyond recording the waypoints for the selected point and structure edge, though it does introduce additional elements of subjectivity into the weight

measurements. Ignoring the weights completely would introduce bias, as it would only be approximately unbiased if dwellings were identical in size and equidistant.

In addition to the above concerns weight calculations, another potential issue with this group of methods is that there are points in the PSU that would not lead to the selection of any households. Consider the shaded area of figure 5. If any of these points were selected, the interviewer would not find any household before she left the boundaries of the PSU. This issue raises questions for the field protocols. Should interviewers stop at the PSU boundary, or should they continue and select housing units outside of the selected PSU? If the former, how would the interviewer know where the PSU boundaries are? If the latter, the probabilities should be adjusted for the fact that the A_i region extends outside of the PSU, which is not straightforward. Additional structures outside of the boundaries of the PSU would need to be mapped, requiring additional preparation time. For the purposes of this paper, we mapped all households in a 50 meter buffer zone around the PSU boundaries. This increased the number of structures required from 309 to 408, 68 to 207, and 353 to 724, respectively, nearly doubling the required mapping time if manual delineation is used. A third option would be to allow interviewers to travel outside of the PSU in search of a selected household, but then remove these interviewed households outside the selected PSUs from the data set, because their probabilities of selection are too complex to calculate. This approach preserves the probabilities of selection and is easy for the interviewer to implement, but deleting data is inefficient in terms of cost.

3.5 Random Walk

There are many different implementations of the random walk procedure, of which each invokes choosing a starting point within the selected area and then proceeding along a path, selecting every k^{th} household. The methods differ in how the path is defined. In this paper, we follow the method used by the Afrobarometer survey. The walking instructions are:

“Starting as near as possible to the SSP [Sampling Start Point], the FS [Field Supervisor] should choose any random point (like a street corner, a school, or a water source) being careful to randomly rotate the choice of such landmarks. From this point, the four Fieldworkers follow this Walk Pattern: Fieldworker 1 walks towards the sun, Fieldworker 2 away from the sun, Fieldworker 3 at right angles to Fieldworker 1, Fieldworker 4 in the opposite direction from

Figure 5: Area which does not lead to selection of household

Source: Authors' diagram based on PSU boundaries and Google Earth images

Fieldworker 3.... Walking in their designated direction away from the SSP, they will select the fifth household for their first interview, counting houses on both the right and the left (and starting with those on the right if they are opposite each other). Once they leave their first interview, they will continue on in the same direction, and select the tenth household (i.e., counting off an interval of ten more households), again counting houses on both the right and the left. If the settlement comes to an end and there are no more houses, the Fieldworker should turn at right angles to the right and keep walking, continuing to count until finding the tenth dwelling” (Afrobarometer, pg. 35).

To simulate the random walk in the Mogadishu context, we replicate the Afrobarometer protocols to the extent possible. First we selected a random starting point (since it is not possible to identify landmarks with the level of detail available on the maps, we simply use a random point as the path start). To simulate the direction of the sun, a random angle is chosen and the direction of the interviewer’s path assigned at 90 degree intervals. For example, if 13 degrees from due north was selected, then the four paths would be at 13 degrees, 103 degrees, 193 degrees, and 283 degrees. From these lines, it was assumed that every dwelling within five meters on either side of the direction of walking was within the interviewer’s line of sight. These dwellings were sequentially numbered and every fifth dwelling selected. If the interviewer reached the PSU boundary before selecting the requisite number of households, the path made a 90 degree turn and continued. If each of the four interviewers selected three households, the total cluster size would be twelve. In order to ensure comparability with the other methods, each of which aimed to select ten households, we dropped the last two selected households.⁴

4. Results

4.1 Simulations

For each of the sampling methods discussed above and the three different methods of allocating consumption values to households (random, some spatial clustering, extreme clustering), we simulated 10,000 samples and calculated the mean for each one. We report in table A2 in the appendix the mean, standard deviation,⁵ 5th percentile, and 95th percentile of the distribution across all 10,000 samples and evaluate the different sampling approaches in terms of their bias and variance. If a sampling method is unbiased, the expected value of the sample means should be 40, by design the true mean consumption in each simulated PSU.

While generally it was possible to implement all of the methods in our simulations, there were notable challenges with three of the designs. In simulating the Qibla method, certain selected points did not lead to a selection within the EA. The impact was largely negligible in Heliwa or Hodon, where only 0.4 percent and 1.4 percent, respectively, of the total area led to no selection, but in Dharkinley, the smallest and most regular of the PSUs tested, 13 percent of the area led to no selection within the PSU, substantially decreasing the efficiency of that method. Then in the implementation the grid selection method, there was little control over the number of households in each grid square. In some cases, grid squares were

⁴ The analogous action in the field would be for the supervisor to rotate the additional interviews between interviewers to assure an even workload, though most likely in the design stage the cluster size would have been set to be evenly divisible among interviewers.

⁵ The standard deviation of the distribution is the standard error of the estimate of the mean.

empty or did not have the minimum number of structures to achieve the expected sample size. In the most extreme case of the large and sparsely populated PSU of Heliwa, when 50 x 50 m grid squares were used, 42 of the 169 grid squares contained no structures. Of those remaining, a further 90 had less than the necessary five structures. Therefore the grid squares were combined into 100 x 100 m squares. After combination, there were 51 grid squares, 7 of which were empty, but 16 continued to contain less than the minimum number of structures. For the simulations, we dropped grid squares without households, though this would likely not be possible in true field implementation, leading to cost inefficiencies.

Figure 6: Heliwa PSU with 50 meter grid square overlay

Source: Authors' diagram based on PSU boundaries and Google Earth images

Finally, there are several documented problems with random walk methods, as we discussed in Section 2. One difficulty not previously discussed in the literature but encountered in the simulated implementation was that the protocols above fail in certain situations. As shown by figure A7 in the appendix, depending on the start point and direction, it may not be possible to turn right and remain within the boundaries of the PSU. The interviewer would need to violate the protocols or seek advice from a supervisor to continue implementation.

4.2 Bias and Variance

The mean, standard deviation, and coefficient of variation are shown in figures 7 to 10 and in table A2 in the appendix for the eight methods under the three different consumption values, for each PSU as well

as overall.⁶ From this table, we can evaluate how well each method worked in terms of bias and variance. From a true mean of 40, it was unsurprising that the full listing / satellite mapping method showed the most consistently efficient and unbiased results. Segmentation also showed consistently unbiased results but had higher variance for higher degrees of clustering in the underlying distribution due to homogeneity within the segments. The Qibla method with the full weights yielded unbiased results but with wide confidence intervals, though these are likely artificially wide in the simulations. (See note in the technical appendix for more detail.) In addition, the wide confidence intervals are partially driven by a few outlier values. 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 percent depending on the degree of clustering. There is also evidence of the trade-off between bias and variance. The weights for the proxy methods are based on where the random point is selected, which is necessarily shorter than the full shadow width, truncating the values of the weights. While this introduces a bias into the measures, it also limits the possibility of having large weights for outlier values, which increase the variance. The width of the confidence interval showed almost no impact when clustering was introduced. There was also little difference in terms of bias and variance between proxy weights 1 and proxy weights 2, indicating there is little information lost if the categorical method is implemented correctly. The unweighted version consistently underestimated the true mean, though showed narrow confidence intervals, due to the weighting loss, or the increase in variance resulting from the application of weights (see Eckman and West, forthcoming, and Kish 11.2C, 1965 for further discussion). The final two sampling methods both over-estimate the means with a bias up to ten percent for the clustered distributions. This is most likely due to grid squares which do not have the required number of dwellings, so that the final sample size did not reach 10. The random walk as noted above, is not theoretically unbiased and this is reflected in the simulation results.

Across the three PSUs, there are also some important differences in the methods, as shown in the violin graphs in figure A8 and A9 in the appendix. Dharkinley, despite being the most regular in terms of layout, was problematic for many of the methods. Satellite mapping, segmentation, and random walk all showed a second bulge in the distribution about 20 percent above the true mean, as compared to an expected smooth normal distribution. The full weighting scenarios for the Qibla method also had the most difficulties in Dharkinley, generating a small number of outlier estimates over 1000 compared to a true mean of 40. In contrast in the chaotic Hodon, there were no issues with satellite mapping and the full weight Qibla method estimates were on par in precision with segmentation for the clustered distributions. The Qibla proxy methods, however, showed substantial bias when clustering was introduced with only a slight decrease in variance. Random walk also had substantial difficulties in Hodon, showing both high levels of bias and variance. The bias is caused by the interviewer instructions, which predefine the path an interviewer has to take. Even though the starting location is randomly selected, interviewers tend to reach certain areas with a higher likelihood. Estimates for variables which are correlated with these unevenly distributed selection probabilities are biased.

⁶ “Overall” is defined as a constructed population with three equally weighted strata rather than randomly selected PSUs with their own probabilities of selection.

Figure 7 : Mean and Confidence Intervals - Overall

Figure 8 : Mean and Confidence Intervals – Dharkinley

Figure 9 : Mean and Confidence Intervals - Heliwa

Figure 10 : Mean and Confidence Intervals - Hodon

4.3 Ease of Implementation and Remote Supervision

In conflict and capacity constrained environments, such as Mogadishu, the ease of implementation and options for remote supervision were also necessary considerations in the selection of the final methodology. Satellite mapping requires little specialized training beyond the use of a GPS device for navigation as target households were selected in the office. The Qibla and random walk methods similarly require the ability to navigate with a GPS to the selected point, but also require additional training for interviewers to correctly implement household selection protocols, which are substantially more complex with random walk. The proxy weights version of the Qibla methods also require interviewers to be training on using GPS for field measurements. Segmentation and grid methods are the most difficult to implement in the field as they require interviewers to identify the boundaries of sub-sections, which in the case of the grid method may not follow landmarks and may cross through structures.

For the purposes of remote verification, the two main GPS-based tools available to for supervision are waypoints and tracks. Waypoints record the latitude and longitude coordinates of a given location while track records the path of the GPS from the time it was activated. The satellite mapping can be effectively supervised remotely with waypoints. The point recorded by the interviewer when they arrive at the household can be compared to the coordinates of the selected household to ensure they correspond to the same structure. This would be most effective in sparsely populated spaces with little overhead obstruction. Verification would be more difficult in dense urban areas where the minimum of ± 15 feet (5 meters) accuracy of the GPS could lead to multiple possible structures, or if heavy overhead cover of metal roofs blocks GPS signals. The Qibla and random walk methods would both use a way point to identify the starting point then the track to confirm the path taken. Grid and segmentation can both use waypoints to confirm points are within selected areas, and it may also be possible to use tracks to confirm the listing process if strict protocols are used (ie. Start in the NE corner and continue clockwise) though the intersection of the interviewers paths may make results less clear.

4.4 Replacements

Due to high transportation costs, most surveys in the developing world use replacements for non-response due to refusals or out of sample selections. This is done either through selecting additional households from the PSU listing exercise, as is recommended in the World Bank's Living Standards Measurement Study (Grosh and Munoz, 1996), or selecting a neighboring structure based on field protocols, such as selecting the dwelling immediately to the right (Lowther et al, 2009). While replacements for out of sample selections with new random points does not introduce bias, it is inefficient and increases costs. For non-response due to refusal, it is likely to be non-random, and therefore replacements will create at least some degree of bias in the data. The reason and method for the replacement may influence the degree. If refusals tend to come from households in the highest and lowest wealth quintiles, as the opportunity cost of their time is high, and replacements come from the main part of the distribution, the use of replacements will attenuate the variation in the sample. This may cause the results to underestimate measures such as inequality that depend on accurately capturing the extremes of the distribution. When using a replacement method that uses near neighbors, if structures

are abandoned or commercial buildings, those households living adjacent may be systematically different from the remainder of the PSU. In addition, those households near the boundary of the PSU would have a lower probability of selection since there are fewer households near them that would lead to them being selected as replacements.

Of the methods discussed above, segmenting and gridding require a short listing exercise at which time non-eligible structures can be excluded. Satellite mapping and the Qibla method rely on maps that cannot differentiate based on eligibility, and are therefore more vulnerable to issues with out of sample selections. In addition, regardless of method, the survey protocols should address procedures for the inevitable refusals, which may be more likely in conflict areas.

5. Discussion

Ultimately the most appropriate method for second stage sampling in any survey depends on a trade-off between cost, necessary precision, and tolerable bias. In conflict zones, these decisions are further complicated by time pressures, available back office resources, and security concerns. Satellite mapping, segmentation, and the Qibla methods with full area weights are all probability methods for which it is possible (though necessarily not easy) to calculate weights, and thus all produce unbiased estimates of the population mean. Of these options, the simulations demonstrated that satellite mapping yielded the most consistently unbiased and efficient design, under the assumption that recent maps are available and potential issues with out-of-sample buildings can be adequately addressed. The Qibla method provides promising results in the simulations but has yet to be tried in the field. The proxy weight variations of the Qibla method also show promise as they remove the requirement of updated satellite maps and greatly reduce the calculation burden for the weights, but do show substantial bias in certain circumstances. The non-probability methods, random walk and the unweighted Qibla method, do not produce unbiased results. Random walk, in particular, did not perform well in the simulations despite being common practice for many surveys.

The simulations also showed the implications of bias in the estimates can be substantial in terms of policy conclusions drawn from the data. In this study, the main indicator was household consumption, which underpins poverty calculations in much of the developing world. For a hypothetical poverty line set at the bottom 40 percent of the population, the bias resulting from using a random walk over satellite mapping would lead to an under-estimation of a poverty rate by five percentage points. Given the expanding availability of satellite maps and decreasing costs of GPS technology, much of which is integrated into the phones and tablets used by interviewers, alternative methods based on probability sampling may reduce bias with little impact on cost or complexity of implementation.

Beyond the simulated results, a number of questions remain that can only be addressed by field testing. For example, it is not possible to discuss the cost considerations of the choice of method nor to comparatively discuss the implications on interviewer safety. Also, the simulations assume perfect implementation and further research is needed on the implications of human error or of outdated maps.

In the case specifically discussed here, the Mogadishu High Frequency Survey Pilot, the team opted to use segmentation as a compromise between preparation time, ease of implementation, and the time and complexity necessary for the weight calculations. The implementation was generally successful despite a number of difficulties in the field. Teams occasionally encountered high-level security threats and exploitative rent-seeking from local leadership. The complexity of the survey protocols, including the sampling design, slowed the implementation of the survey. Also a substantial number of observations had to be discarded because the interviewed points did not fall within the boundaries of the selected segments or because interviewed households did not appear on segment listing forms. Regardless of these challenges, however, it was possible to implement a complex and yet rapid, high-quality survey in one of the most challenging urban contexts known to date.

References

- Afrobarometer Network, 2014. Afrobarometer Round 6 Survey Manual.
- Afzal, Marium, Jonathan Hersh, and David Newhouse. 2015. "Building a Better Model: Variable Selection to Predict Poverty in Pakistan and Sri Lanka." Mimeo.
- Alt, C., Bien, W., Krebs, D., 1991. Wie zuverlässig ist die Verwirklichung von Stichprobenverfahren? Random route versus Einwohnermeldeamtsstichprobe. ZUMA-Nachrichten 28, 65–72.
- Aminipouri, M., Sliuzas, R., Kuffer, M., 2009. Object-oriented analysis of very high resolution orthophotos for estimating the population of slum areas, case of Dar-Es-Salaam, Tanzania, in: Proc. ISPRS XXXVIII Conf. pp. 1–6.
- Barry, M., Rüther, H., 2001. Data collection and management for informal settlement upgrades, in: Proc. International Conference on Spatial Information for Sustainable Development. Citeseer.
- Bauer, J.J., 2014. Selection Errors of Random Route Samples. Sociological Methods & Research 0049124114521150. doi:10.1177/0049124114521150
- Bennett, A., Radalowicz, A., Vella, V., Tomkins, A., 1994. A Computer Simulation of Household Sampling Schemes for Health Surveys in Developing Countries. International Journal of Epidemiology 23,6: 1282-1291.
- Bien, W., Bender, D., & Krebs, D. 1997. DJI-Familiensurvey: Der Zwang, mit unterschiedlichen Stichproben zu leben. In *Stichproben in der Umfragepraxis* pp. 127–147. Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-322-86533-5_10
- Blohm, M. 2006. Data Quality in Nationwide Face-to-face Social Surveys. Retrieved from <http://www.ons.gov.uk/ons/about-ons/get-involved/events/events/q2006---european-conference-on-quality-in-survey-statistics-24-26-april-2006/agenda/session-17-wednesday.pdf>
- Dreiling, K., Trushenski, S., Kayongo-Male, D., & Specker, B. 2009. Comparing household listing techniques in a rural Midwestern vanguard center of the national children's study. Public Health Nursing, 26(2), 192-201.
- Eckman, S., Himelein, K., Dever, J., forthcoming. New Ideas in Sampling for Surveys in the Developing World, in: Johnson, T.P., Pennell, B.-E., Stoop, I., Dorer, B. (Eds.), *Advances in Comparative Survey Methodology*. 3MC.
- Eckman, S. & West, B. forthcoming. Analysis of Data from Stratified and Clustered Surveys, in Wolf, C., Joye, D., Smith, T., and Fu, Y. (Eds.), *Handbook of Survey Methodology*. Sage.
- Eckman, S. and Koch, A. 2016. "Are High Response Rates Good for Data Quality? Evidence from the European Social Survey" Paper under review.
- Gallup, 2014. Farm workers pessimistic about their lives [WWW Document]. URL <http://www.gallup.com/poll/169019/farm-workers-africa-pessimistic-lives.aspx> (accessed 1.27.15).
- Gallup, 2015. World Poll Methodology [WWW Document]. URL <http://www.gallup.com/poll/105226/world-poll-methodology.aspx> (accessed 1.27.15).
- Grais, R.F., Rose, A.M., Guthmann, J.-P., 2007. Don't spin the pen: two alternative methods for second-stage sampling in urban cluster surveys. Emerging Themes in Epidemiology 4, 8. doi:10.1186/1742-7622-4-8
- Grosh, M.E., Munoz, J., 1996. A manual for planning and implementing the living standards measurement study survey (No. LSM126). The World Bank.
- Harter, R., Eckman, S., English, N., O'Muircheartaigh, C., 2010. Applied sampling for large-scale multi-stage area probability designs. Handbook of survey research 2, 169–199.
- Himelein, K., Eckman, S., Murray, S., 2014. Sampling Nomads: A New Technique for Remote, Hard-to-Reach, and Mobile Populations. Journal of Official Statistics 30.
- Hoffmeyer-Zlotnik, J. H. 2003. New sampling designs and the quality of data. Developments in Applied Statistics. Ljubljana: FDV Methodoloski Zvezki, pp. 205–217.

- Kish, L. (1965). *Survey Sampling*. New York: Wiley.
- Kolbe, A.R., Hutson, R.A. 2006. Human Rights Abuse and Other Criminal Violations in Port-Au-Prince, Haiti: A Random Survey of Households. *The Lancet* 368 (9538): 864–73. doi:10.1016/S0140-6736(06)69211-8.
- Kondo, M.C., Bream, K.D.W., Barg, F.K. Branas, C.C. 2014. A Random Spatial Sampling Method in a Rural Developing Nation. *BMC Public Health* 14 (1): 338.
- Kumar, Naresh. 2007. "Spatial Sampling Design for a Demographic and Health Survey." *Population Research and Policy Review* 26 (5-6): 581–99.
- Lowther, S.A., Curriero, F.C., Shields, T., Ahmed, S., Monze, M., Moss, W.J., 2009. Feasibility of satellite image-based sampling for a health survey among urban townships of Lusaka, Zambia. *Tropical Medicine & International Health* 14, 70–78. doi:10.1111/j.1365-3156.2008.02185.x
- Mneimneh, Z.N., Axinn, W.G., Ghimire, D., Cibelli, K.L., Alkaisy, M.S., 2014. Conducting surveys in areas of armed conflict, in: *Hard-to-Survey Populations*. Cambridge University Press.
- Pape, U. and Mistiaen, J. 2015. Measuring Household Consumption and Poverty in 60 minutes: The Mogadishu High Frequency Survey. World Bank / Proceedings of the Annual Bank Conference on Africa. Berkeley, CA.
- Särdnal, C.E., Swensson, B., Wretman, J.H., 1992. *Model assisted survey sampling*. Springer.
- Turkstra, J., Raithelhuber, M., 2004. *Urban Slum Monitoring*.

Appendix

Table A1 : Description of sample PSUs							
Location	Total PSU Area (m ²)	Total PSU + Buffer Area (m ²)	Area in which no households would be selected with Qibla method (% of total)	Number of Structures	Number of Structures (including buffer)	Imagery date	
Hodon	42,615	95,707	0.4%	309	408	March 14, 2014	March 13, 2015
Dharkinley	24,390	65,447	1.4%	68	207	December 25, 2013	March 13, 2015
Heliwa	345,157	477,252	13.0%	353	724	March 14, 2014	March 10, 2015

Table A2: Main Results

Method/Clustering	Dharkinley (EA4)				Heliwa (EA5)				Hodon (EA6)				Overall			
	<i>mean</i>	<i>sd</i>	<i>p5</i>	<i>p95</i>	<i>mean</i>	<i>sd</i>	<i>p5</i>	<i>p95</i>	<i>mean</i>	<i>sd</i>	<i>p5</i>	<i>p95</i>	<i>mean</i>	<i>sd</i>	<i>p5</i>	<i>p95</i>
True Mean	40.0				40.2				40.0				40.0			
Full Listing / Satellite Mapping																
Randomly assigned	40.1	11.6	26.2	60.7	39.8	9.4	27.0	58.3	40.1	9.5	27.1	58.2	40.0	6.2	30.9	51.2
Some spatial clustering	39.8	11.6	25.8	60.7	40.0	9.6	27.3	58.9	39.9	9.6	26.9	58.2	39.9	6.3	30.9	51.5
Extreme clustering	39.9	11.6	25.9	60.6	40.0	9.6	27.3	58.6	39.9	9.5	27.1	58.3	39.9	6.2	30.9	51.5
Mecca method (accurate weights)																
Randomly assigned	40.2	29.8	19.0	104.3	40.5	30.1	17.8	82.1	40.3	18.7	18.4	76.7	40.3	15.4	24.5	70.0
Some spatial clustering	39.8	45.2	19.8	65.7	40.3	34.8	15.2	94.6	40.2	17.0	21.5	69.3	40.1	19.8	24.7	62.9
Extreme clustering	40.0	45.3	19.9	66.0	40.6	33.4	15.1	97.5	40.3	18.2	21.7	66.8	40.3	19.6	24.8	64.1
Mecca method (proxy weights 1)																
Randomly assigned	37.6	18.0	20.4	67.6	39.2	18.6	18.9	68.9	39.7	15.2	22.3	64.4	39.3	11.0	25.8	58.2
Some spatial clustering	37.7	16.7	21.3	66.3	38.2	19.2	18.2	69.0	47.1	15.8	27.4	73.5	42.5	11.3	28.4	61.9
Extreme clustering	35.9	13.7	20.8	55.2	38.2	19.5	18.1	69.2	47.6	15.3	27.5	73.0	42.6	11.2	28.5	61.7
Mecca method (proxy weights 2)																
Randomly assigned	37.3	17.0	20.4	64.9	39.5	18.6	18.9	69.2	40.0	15.8	22.1	65.3	39.5	11.2	25.9	58.5
Some spatial clustering	37.3	16.2	21.2	64.7	38.2	19.0	18.1	68.7	47.2	16.2	27.0	74.1	42.5	11.4	28.2	61.9
Extreme clustering	35.4	13.2	20.6	54.3	38.2	19.6	18.0	69.1	47.6	16.2	27.2	73.4	42.5	11.5	28.3	62.0
Mecca method (no weights)																
Randomly assigned	34.4	8.3	24.0	52.3	38.9	8.3	26.8	53.6	40.4	9.9	27.2	59.4	38.0	5.2	30.5	47.5
Some spatial clustering	35.6	8.6	24.8	54.4	34.2	7.1	24.0	46.6	44.0	8.5	32.0	59.5	38.0	4.6	31.2	46.3
Extreme clustering	36.1	8.9	24.9	55.5	34.4	8.2	23.8	50.1	44.6	8.6	32.5	60.3	38.4	4.9	31.3	47.6
Segmentation																
Randomly assigned	39.8	11.2	24.7	60.2	39.9	9.6	26.8	58.8	39.9	9.3	27.2	58.1	39.9	6.2	30.9	51.2
Some spatial clustering	40.5	18.5	18.4	78.2	39.8	16.7	16.9	71.5	39.6	17.2	16.6	71.5	39.8	11.1	23.4	59.9
Extreme clustering	40.6	17.6	18.6	69.0	39.8	19.0	16.8	78.4	39.8	19.1	16.3	78.0	39.8	12.4	23.1	63.4
Grid																
Randomly assigned	40.8	14.8	23.1	70.8	39.8	11.0	25.9	60.5	41.1	11.5	25.9	62.5	40.5	7.4	30.2	54.3
Some spatial clustering	38.5	20.9	15.6	81.4	41.4	19.0	16.3	78.2	45.7	20.1	19.0	81.3	43.2	12.8	24.8	66.1
Extreme clustering	39.0	25.0	15.6	97.5	42.9	24.7	16.4	83.2	45.6	23.0	18.4	90.3	43.9	15.4	23.7	72.6
Random walk																
Randomly assigned	39.3	12.0	25.5	62.0	40.5	10.1	27.2	60.4	39.9	9.8	26.8	59.3	40.1	6.5	30.8	52.1
Some spatial clustering	38.2	10.1	27.1	56.7	40.2	11.3	24.6	62.0	45.6	18.1	18.3	77.8	42.6	10.0	27.3	60.1
Extreme clustering	38.7	10.5	27.0	57.3	39.2	10.2	24.5	58.7	46.2	20.9	17.5	87.3	42.5	11.0	26.7	63.4

Figure A1 : Dharkinley

Figure A2 : Heliwa

Figure A3 : Hodon

Figure A4: Example of “shadows” for Qibla method

Examples of path of random walk

Figure A8 and A9: Violin graphs of simulated values

Technical Appendix

In the Qibla method households could be selected multiple times, if more than one randomly selected point fell into their selection region. In such a case, the formula to estimate the expected values of variable X is:

$$E(X) = \frac{\sum_{i=1}^n x_i \cdot w_i}{\sum_{i=1}^n w_i} \quad (\text{A1})$$

where x_i is the value of X of person i ; $w_i = \frac{1}{p_i}$ is the weight of person i defined as the inverse of the 1-draw selection probability (p_i); and n is the number of draws in the sample (not the number of unique cases drawn). Equation A1 gives an unbiased estimate of the population mean of X when n is large. In particular, the denominator converges to the sample size multiplied by the population size N as the sample gets larger:

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n w_i = n \cdot N. \quad (\text{A2})$$

However, in our simulated application of the Qibla method, we make only ten selections in each EA. With such a small sample size, $\sum_{i=1}^n w_i$ is not necessarily close to $n \cdot N$ and leads to bias in estimate of the mean of X . Given that we know N we use $n \cdot N$ as a replacement for the denominator in Equation A1:

$$E(X) = \frac{\sum_{i=1}^n x_i \cdot w_i}{n \cdot N} \quad (\text{A3})$$

which correctly estimates the expected values for X , but has a higher variance than the standard formula.

In large surveys which sample from multiple EAs, the total n is much larger than in our simulation. Therefore, the standard weighting approach will lead to a correct estimation of expected values and smaller variances than in the shown simulation. It is only in this paper, where we look at EA-level means, that n is small and Equation A1 fails and we must use Equation A3; in most normal applications of the Qibla method, where country or region level means are of interest, Equation A1 would be appropriate.