A census-independent sampling strategy for a household survey in Myanmar

Juan Muñoz¹ and Wijnand Langeraar²

November 2013

Table of Contents

1. Introduction and background ................................................................................................................. 2
2. The proposed sampling strategy ............................................................................................................. 3
3. The LandScan database as a first-stage sample frame ........................................................................... 3
   3.1. The LandScan database .................................................................................................................. 3
   3.2. Administrative stratification .......................................................................................................... 5
   3.3. Urban/Rural stratification .............................................................................................................. 6
   3.4. Selection of Primary and Secondary Sampling Units ..................................................................... 6
4. Segmentation of highly populated areas and development of the tertiary stage sample frame .... 7
   4.1. Detailed description with GIS aspects ............................................................................................. 7
   4.2. Selection of Tertiary Sampling Units ............................................................................................. 15
5. Household listing and household selection .......................................................................................... 16
   5.1. Household Listing .......................................................................................................................... 16
   5.2. GPS Support to the household listing operation ............................................................................. 16

¹ Sistemas Integrales. 🌐 juan.munoz@ariel.cl ✉ Casilla 13168, Santiago, Chile ☎ +562-2639-4554 (x 156)
² Sistemas Integrales. 🌐 winedl@gmail.com ✉ Jalan Pengembak 30-C, 80228 Sanur, Bali, Indonesia ☎ +62-818-351-314
**1. Introduction and background**

Most national household surveys visit a sample of from a few to several thousands of households, selected with known (albeit often unequal) probabilities from among all households in the country. The total sample is typically allocated into geographical strata in order to deliver estimates for the major administrative subdivisions of the country. Within each stratum, the sample is selected in two stages, with small area units of a few hundred households (called Primary Sampling Units, or PSUs) in the first stage, and individual households in the second stage. The first-stage sample frame (i.e., the list of all PSUs in the country) is based on the information collected by the most recent census, and thus tends to be common to many surveys, whereas the second-stage frame tends to be specifically developed for each survey, by way of a household listing operation conducted in the selected PSUs. PSUs are typically selected with probability proportional to size (the so called pps technique), and then a fixed number of households (called a cluster) is selected with equal probability in each selected PSU.

This sampling strategy – hereafter referred to as “the standard strategy” – is critically dependent on reliable and accessible census databases and census cartography. National statistical agencies are the sole source of such data in many countries, which can make them hard to reach. Even when available, census databases are often unreliable and almost always outdated. The conditions are particularly acute in Myanmar, where a reliable census has not been conducted for decades and household surveys are urgently required.

In this note we propose an alternative strategy, intended to deliver a sample of households with well-defined probabilities, on the basis of Geographic Information System (GIS) techniques, GPS, high-resolution satellite imagery, and additional information sources other than censuses, in particular the LandScan population database.

The following GIS software packages have been used in the development of this alternative strategy:

- Google® Earth Prov. 7.1.1.1888
We specifically focus on Myanmar, but the techniques can be adapted to other countries.

2. The proposed sampling strategy

We use as our basic sample frame the list of 840,929 LandScan cells for Myanmar. Each cell is a rectangle of 30 x 30 arc-seconds (approximately 900 x 900 meters), for each of which the database reports an estimated population. We assume that the sample will use the 14 States – Myanmar’s major administrative units – as strata.

Within each State, we intend to implement a 4-stage sampling strategy, where

- the first stage units are 3x3 super-cells (i.e., rectangles of approximately 2,700 x 2,700 meters, each composed of nine of the original LandScan cells);
- the second stage units are individual cells (or groups of cells);
- the third stage units are territorial segments with recognizable boundaries; and
- the fourth stage units are individual households.

The first two stages are fully computerized. The third stage requires human intervention, but is conducted in the office, on the basis of available satellite imagery. The fourth stage is a field operation, essentially equivalent to the final stage of the standard strategy.

The first- and second-stage sampling units will be selected with pps, using as a measure of size the population of the cells, as reported in the LandScan database. The third-stage units will also be selected with pps, using a measure of size a quick count of the number of dwellings in each segment, registered as by-product of the segmentation process. In the final stage, a fixed number of households (a cluster) will be chosen in each of the selected segments by systematic equal probability sampling (sepsam), just as in the standard sampling strategy. The result will be a nearly self-weighted sample in each stratum.

The following sections give the details and the rationale behind each of the proposed stages.

3. The LandScan database as a first-stage sample frame

3.1. The LandScan database

LandScan is a global population database developed by the Oak Ridge National Laboratory (ORNL), US Department of Energy. The development is being funded by the US Department of Defense. Worldwide population data is presented in an ESRI compatible raster format of 30 by 30 arc seconds (0.00833 decimal degrees) and is output in a geographical coordinate system with longitude and latitude as x and y map units, and with World Geodetic System (WGS) 1984 as map datum. The values of the cells are integer population counts representing a 24-hour average, or ambient, population distribution.
The LandScan database is available free for US Government agencies. All others do also have access to the data but have to pay for it. Data can be purchased for each country separately, region, or for the entire world in one go.³ We purchased the latest (2012) LandScan database for Myanmar (840,929 raster cells) and used it as a test bed for the development of our sampling strategy. The examples and images presented below are mostly based on that experience.

Since our first two sampling stages are to be implemented with pps, using the LandScan population figures as measure of size, a natural question is whether these figures are accurate enough for these purposes. Before answering that question, we should point out that our measure of size does not need to be exact. In order to develop unbiased estimates from the sample, the selection probabilities only need to be positive and known – which will certainly be the case. Having said this, we also believe that the LandScan figures are very well correlated with the “real” number of households, meaning that that our estimators will not be just unbiased, but also very efficient.

ORNL claims that LandScan is the finest resolution global population distribution data available. The LandScan algorithm uses spatial data and imagery analysis technologies and a multi-variable dasymetric⁴ modelling approach to disaggregate census counts within an administrative boundary. Much of the spatial data and imagery processed in the LandScan model is provided by the National Geospatial-Intelligence Agency (NGA). Ancillary information is coming a.o. from the Census Bureau, US Department of Agriculture, NASA. Data used in the determination of population per grid cell include (i) census and city population information, (ii) various global vector data sets such as VMAP, (iii) digital terrain models like the ones derived from the Space Shuttle Radar Topographic Mission (SRTM), (iv) land cover data resulting from Landsat TM and other sources, (v) various high-resolution earth observation satellite imagery such as IKONOS, QuickBird, etc. We have independently assessed the quality of the LandScan figures through visual comparisons with the other sources of information at our disposal, including Google Earth imagery, both for low- and high-density areas.⁵

The examples and images presented below are thus based on the Myanmar database. The LandScan data set is in raster format. Each of the 840,929 cells for Myanmar has a unique ID number which we have adopted in order to identify and distinguish the particular cells. The numbering is from upper left (NW) to lower right (SE).

The cell surface is half a nautical mile square at the equator and varies proportionally with the cosine of the latitude. We thought it practical to keep the GIS work unprojected, with geographical coordinates as the x-y system, because the cell area is irrelevant: only the population count is important. In a geographical coordinate system the cells conveniently appear as squares at any latitude.

Figure 1 below is a color-coded representation of the 2012 LandScan dataset:

³ LandScan data can be purchased from Scott Edwards, LandScan Product Manager – East View Information Services – 10601 Wayzata Blvd, Minneapolis, MN 55305 - Phone: +1.952.252.1201 | Fax: +1.952.252.1202 – scott.edwards@eastview.com | www.eastview.com.

⁴ Dasymetric mapping depicts quantitative areal data using boundaries that divide the area into zones of relative homogeneity with the purpose of better portraying the population distribution.

⁵ For more information on LandScan, see http://web.ornl.gov/sci/landscan/landscan_documentation.shtml
The above picture shows the central part of the country with LandScan population data values rendered with a quantile classification in 13 classes (each class contains an equal number of cells - blue is sparse and red is dense population). With this rendering, corridors of higher population densities along major infrastructure become pronounced. SRTM\textsuperscript{6} shaded relief is transparently overlain to demonstrate correlation of population with relief. Also shown are State boundaries.

### 3.2. Administrative stratification

The LandScan database does not explicitly report the administrative affiliation of the cells, but we were able to add this information based on the State, District and Township polygons available to us from other sources. A few of cells can overlap two administrative units or more, but most of these “boundary” cells are unpopulated so this does not represent a problem. Only the affiliation of cells to States is likely to be needed for explicit stratification. The affiliation to Districts and Township will almost certainly be used for implicit stratification only.

\footnote{6 The Shuttle Radar Topography Mission (SRTM) is an international project spearheaded by NGA and NASA which obtained digital elevation models on a near-global scale. SRTM flew on board the Space Shuttle Endeavour during an 11-day mission in February 2000. The elevation models derived from SRTM data are used in GIS, and can be downloaded freely over the Internet.}
3.3. Urban/Rural stratification

Sorting the LandScan cells into the officially defined urban and rural categories may be more challenging, since we have no polygons for these. This situation can be dealt with in three possible ways:

- The first option is to do nothing, and simply avoid using the Urban/Rural categories as explicit sampling strata (although the nature of each selected unit can be of course reported by the survey and used for analytic and tabulation purposes). This would be by far the most convenient option, if there are no analytical imperatives to oversample either of the categories.

- The second option is to use an alternative classification that is correlated, but may not be identical, to the official one. We could, for instance, call the cells with 500 people or more “dense”, and those with less than 500 “sparse”, and stratify on this basis.

- The third option is to use the polygons of the areas officially classified as urban, if they do exist and can be obtained easily.

3.4. Selection of Primary and Secondary Sampling Units

The main reason for defining 3x3 super-cells as Primary Sampling Units (PSUs) and individual LandScan cells as Secondary Sampling Units (SSUs), rather than cells as PSUs directly, is the existence of many cells with less than 100 persons (about 20 households). According to the LandScan database, about 15.5% of the population of Myanmar lives in such cells, meaning that, if they were used as PSUs, [a] many of them would be chosen, and [b] The subsequent processes of segmentation, household listing, and cluster selection would be problematic with so few households.

Grouping the original cells into 3x3 super-cells brings the problem down to a manageable size, since only 2.4% of the population lives in super-cells with less than a 100 persons. Most such super-cells are in remote areas and can be dealt with by either

- explicitly excluding them from the scope of the survey, or

- accepting their presence, and the implied burden of conducting the survey in those of them that are selected in the sample.

The need for a choice is not unique to our proposed sampling strategy. Survey practitioners always face a similar dilemma with the standard strategy – and typically opt for the first alternative, although the decision is not always as explicit and well documented as it should be. Once a decision is made, the mechanics for the pps selection of PSUs can proceed in the familiar way.7

The next step is the formation of SSUs (individual cells, or group of cells), within the selected super-cells. SSUs are formed by way of merging any cells with a population of less than 100 to the largest laterally adjacent cell. Figure 2 below shows an example.

---

If the super-cell shown on the left (a part of central Mandalay) were selected in the first stage, then Cell 344355, on the North-West corner, would need to be merged to Cell 344356. No other merges would be required in this case, so the PSUs ends up divided into eight SSUs, as shown on the right.

The second sampling stage – selecting one of the eight SSUs with pps can then proceed as usual.

4. Segmentation of highly populated areas and development of the tertiary stage sample frame

4.1. Detailed description with GIS aspects

LandScan raster cells with dense population will need to be divided into segments, in a process similar to census enumeration area mapping. The segments should have roughly 100 households each, and – more important – the boundaries must be clearly recognizable on the ground.

We converted the LandScan raster into vector formats (points, polygons) so that the attribute data can be more readily used in statistical calculations. For instance, the center and four corner points (5 points per cell in total) were generated with their latitude/longitude and crossed with administrative polygons. Most of the cells will have all five points falling in one State, but the cells straddling the borders have not all of the points in one State. Each cell could be “democratically” allocated to belong to a State with the majority of the five points falling inside. The corner points of the cells also served as control points in the geo-referencing process of Google Earth images.

An important step was to convert the cells to polygons (which appear as squares on the unprojected map). The polygon of a selected cell can eminently be used as the starting point of segmentation digitizing.

The image information from which boundaries are digitized comes from Google Earth Pro. Google Earth tends to have recent and high-resolution images of the more important cities. This is fortunate because the primary selection is done with a probability proportional to size (in our case population) so there is a tendency of selection in high population areas. This means that in case a cell needs segmentation the chance is high to find good-quality Google Earth data for that cell. Individual dwellings can be easily discerned on these city images.
The Google Earth images are downloaded in high resolution (4800 by 4800 pixels) and precision-geo-referenced using the cell corner point coordinates. The polygon of the cell to be segmented further is overlain and the segments can be digitized using a process called heads-up digitizing on-screen.

The attribute table of the polygons contains the information needed to identify the cell (its ‘grid_code’ which is the unique cell ID, and the coordinates of the cell center), the LandScan population figure (‘LSPOP’). We added to those a few more fields to guide in the segmentation process, namely the approximate number of households to expect (APPROX_HH=LSPOP/5) and the number of segments that will need to be drawn (SEGMENT=APPROX_HH/100 + 1). Here is an example of the attribute table of one cell:

<table>
<thead>
<tr>
<th>Shape</th>
<th>grid_code</th>
<th>LSPOP</th>
<th>LATITUDE</th>
<th>LONGITUDE</th>
<th>APPROX_HH</th>
<th>SEGMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polygon</td>
<td>345232</td>
<td>18648</td>
<td>21.995797</td>
<td>96.070676</td>
<td>3730</td>
<td>38</td>
</tr>
</tbody>
</table>

In addition to the LandScan and Google Earth Pro data, the following spatial data have been helpful in the development of the tertiary stage sample frame:

- Latest Landsat-8 imagery;
- The 1 to 50,000-scale topographic digital, geo-referenced map mosaic covering the entire Myanmar land mass, in combination with SRTM-derived shaded relief;
- Open Street Map base map data obtained via ArcGIS Online (handy information because it provides street names);
- Administrative boundaries down to the Township level, derived from the 50k topo map mosaic. This data set is needed for the allocation of LandScan cells to administrative units.

The series of images below show an example of how the above data, linked to the LandScan cells shown as polygons (squares), can be used to perform an actual segmentation in practice. The example is of a group of nine cells (a super-cell) in Mandalay City, the second largest city of Myanmar with about one million inhabitants.

---

8 Available free of charge from the US Geological Department (http://earthexplorer.usgs.gov/).

9 This data set was compiled from scanned hard-copy maps produced in 2006 by the Myanmar Survey department. The 1150 map sheets were scanned, geo-referenced, and mosaicked into nation-wide map coverage. This was done in 2012 at the Myanmar Ministry of Immigration and Population, Department of Population as part of census mapping preparation.
Figure 3 below shows the 345232 super-cell against a background of a very recent Landsat 8 images (taken 23 Sep 2013). The Citadel with the Royal Palace are visible as a slightly tilted rectangle on the right.

Figure 3 – Super-cell 345232 with Landsat 8 taken 23 Sep 2013
Figure 4 covers exactly the same area. The background data come from the 1 to 50,000 scale topographic map mosaic, combined with shaded relief information from SRTM. Thanks to the shading, Mandalay Hill becomes visible just north of the Citadel.

Figure 4 – Super-cell 345232 with 50k scale topomap mosaic
Figure 5 shows the same area, with the OpenStreetMap in the background instead of the topo map mosaic. The SRTM’s Shaded relief is also present.
Figure 6 shows only cell 345232, assumed to have been chosen in the second sampling stage. The cell is shown with the geo-referenced Google Earth Pro image as a backdrop. The population in this cell according to LandScan is 18,648, meaning that it will need to be segmented into about 36 areas of roughly 100 each.

The segmentation is done as follows:

- The polygon cell is saved as a separate shapefile and displayed in ArcMAP (geographic coordinates, map datum WGS 1984)
- The precisely geo-referenced high-resolution Google Earth image is put in the background
- Start an edit session
- Select the cell outline polygon
- Use the Cut Polygon Tool to divide the square into separate segments while making sure that boundaries are chosen that are easily recognizable on the ground. In an urban setting these boundaries in general follow roads and alleys, but also railroads and waterways will do.
Figure 7 shows the Cut Polygon Tool in action. The turquoise polygon is the selected polygon item representing an area that needs to be divided in two. The green line with the dots (the vertices) is snaking through the neighborhood while following a street or alley. The red dot is the vertex where the cursor is at present.

Figure 7 - ArcMap Cut Polygon tool in action
Figure 8 shows the Cut Polygon process completed. The original polygon is now cut in two, along a path which will be recognizable on the ground.

The Cut Polygon digitizing procedure works well because the attribute tables of the newly created polygons can be populated straight away with data (segment number, approximate household quick count). The outline of the original polygon always stays intact. The process is topologically robust because start- and endpoints of dividing digitizing lines will automatically snap to existing border lines, effectively ‘closing’ the polygons.
Figure 9 shows the final output of the segmentation process:

Figure 9 - Cell 345232 with segmentation completed

The numbers inside the segments represent an ‘eyeball’ quick count of the approximate number of roofs inside each segment. This number is used as a measure of size for the next pps sampling stage.

Interesting to note here is that the sum of the approximate number of observed roofs for this cell is 3,850 (with a mean of 107). The expected number of households based on the LandScan population figure was around 3,730. This is not far off the mark!

4.2. Selection of Tertiary Sampling Units

On the basis of the segmentation of the selected SSUs, the selection of Tertiary Sampling Units (TSUs, or segments) can then proceed directly, using the usual pps technique and the quick count of the number of dwellings as a measure of size. We recommend to select and list two segments per PSU (rather than just one), to improve the spread of the sample in urban areas, and to facilitate the estimation of standard errors at the analytical phase of the survey.
5. Household listing and household selection

5.1. Household Listing

In our proposed strategy, the creation of the last-stage sample frame, by way of a full listing of all households in the segment, and the selection of the cluster of households to be visited by the survey from that list, are essentially identical to those of the standard strategy.

We see no reasons for changing this tried and tested approach – there are many risks in trying to implement alternatives based on aerial imagery. Basically, from the air we can only see roofs, but not households. Under a roof we may possibly find a household, but sometimes also many households (in apartment buildings, for instance), and often no households at all (in buildings that are not dwellings).

The survey implementers will face the classical dilemma of either conducting the household listing operation as a separate exercise, or immediately before the interviews in each segment. The first option is a little more expensive, but safer.

5.2. GPS Support to the household listing operation

The function of using GPS is twofold:

- GPS will ensure that the fieldworkers will be able to easily find the Tertiary Sampling Unit, which is either a segment inside a densely populated cell, or an entire sparse cell, and

- GPS will provide support to the household listing operation making sure that all households in the TSU are listed, and that no households out of the TSU are listed.

Imagery of the selected cell areas overlaid with segment polygon outlines will be uploaded into the GPS units. This will effectively replace the necessity of printed maps. However, it may be necessary to print overview maps for each State, with the TSU locations indicated, as shown Figure 10 below.
Figure 10 - Example of an overview map
When the field staff is in process of producing the listing of household in his/her TSU, the location of the GPS receiver will need to be consulted at every household listed, to make sure the household is located within the TSU.

The screen of the GPS receiver would look as in Figure 11 below:

![Simulated GPS screen](image-url)