Sampling Methodology for irregular settlements based on drone images

Javier Ureña, Instituto Tecnológico Autónomo de México
Roberto Sánchez, Instituto Tecnológico Autónomo de México

Irregular settlements are defined by illegal forms of land appropriation according to the legal framework of a specific territory [1]; although, as a phenomena, encompass a wide array of socio-economic and urban characteristics with a particular emphasis on lack of housing quality and urban standards of living [2]. In addition, the absence of urban planning usually results in people locating to high-risk areas, uneven growth and ill-defined street layouts. Since illegality is a common denominator in irregular settlements, many countries fail to provide accurate and updated data, even on the number of households present in a determined area.

The characteristics of irregular settlements hinder the establishment of a well-defined sampling frame when developing survey methodologies. We strive to generalize a framework that can be applied as part of a toolkit to survey households in irregular settlements with the help of drone imaging. Some of the difficulties in establishing a sampling frame arise from the identification of households, our target sampling units. While a form of land census could be used to determine the exact number of households, such methodologies pose scalability problems. Additionally, using drone images to manually count households may prove tricky, since oftentimes it is hard to establish where a household begins and ends just by classifying rooftop materials; such is the case when a household contains two or more rooms with different materials or when many households are contained in high-density areas. As expected, these problems extend to automated image classification.

Requirements

The user of the toolkit must first define different aspects of the survey: the area of interest must already be predominantly an irregular settlement, for it is not the purpose of this methodology to ponder on the legality of the households of interest. Additionally, the user will provide a drone image, cropped to establish a particular territory that is stable over time.

The steps to follow before taking a survey are:

- Definition of the area of interest
- Classification of households
- Extraction of households
- Creation of sampling-frame based on Sample Units (SUs)
- Sampling based on SUs.

Afterwards, the user will proceed to take the survey and analyze the results within the framework of the proposed methodology.

The rationale behind these steps is that while we cannot know the exact number of households and obtain a sampling frame, we can create sampling units of a fixed size in a segmentation process, and sample those units with probability proportional to the expected number of houses in each one (probability proportional to estimated size, or PPES sampling). [*IDEA: será mejor hacerlo por el número esperado de personas que viven en un hexágono? Aunque el objetivo sea obtener información a nivel casa, una casa relativamente grande puede tener albergar varias familias nucleares*] [*Duda: this does not make households mutually exclusive, ie, a house can fall in two or more polygons-- how do we deal with this?*]

Methodology

a) Definition of area of interest

Based on a drone image, the user must define a study area. It may be the whole image or a cropped polygon(s) defined by streets, natural barriers or other criteria that are stable over time. Irregular settlements can change rapidly because of growth or even natural phenomena that alter the landscape, so we encourage to define the area by specific blocks or households.

Notice the objective of this methodology is to survey a particular irregular settlement. If one were to sample a larger area, such as a municipality, via stratified sampling, these irregular settlements would constitute primary sampling units. [*Duda: delimitar el tamaño geográfico de las comunidades con un número específico?*]. This methodology' sampling units will be used a measure to counteract the lack of identifiable single households. [*Nota: otra opción sería hacer

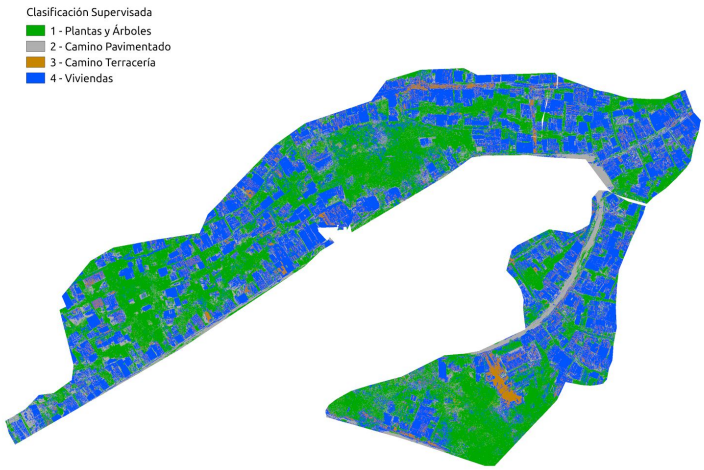
listas de casas dentro de polígonos más grandes específicos, aunque requeriría de una visita extra al asentamiento*].

Example: Cropped drone image of irregular settlement in Arzu, Guatemala.

b) Supervised classification and extraction of households

The following step will examine the distribution of houses in the area of interest. While a manual cut-off of the drone image is possible to obtain the places where households are located, we will use semi-automatic image classification tools to determine the places where households are located.

QGIS' Semi-Automatic Classification plug-in, created by Luca Congedo, is a useful tool for classifying rooftop materials, pavement, dirt and vegetation based on their spectral signature. The objective of this step is to separate households from anything else present in the drone image.

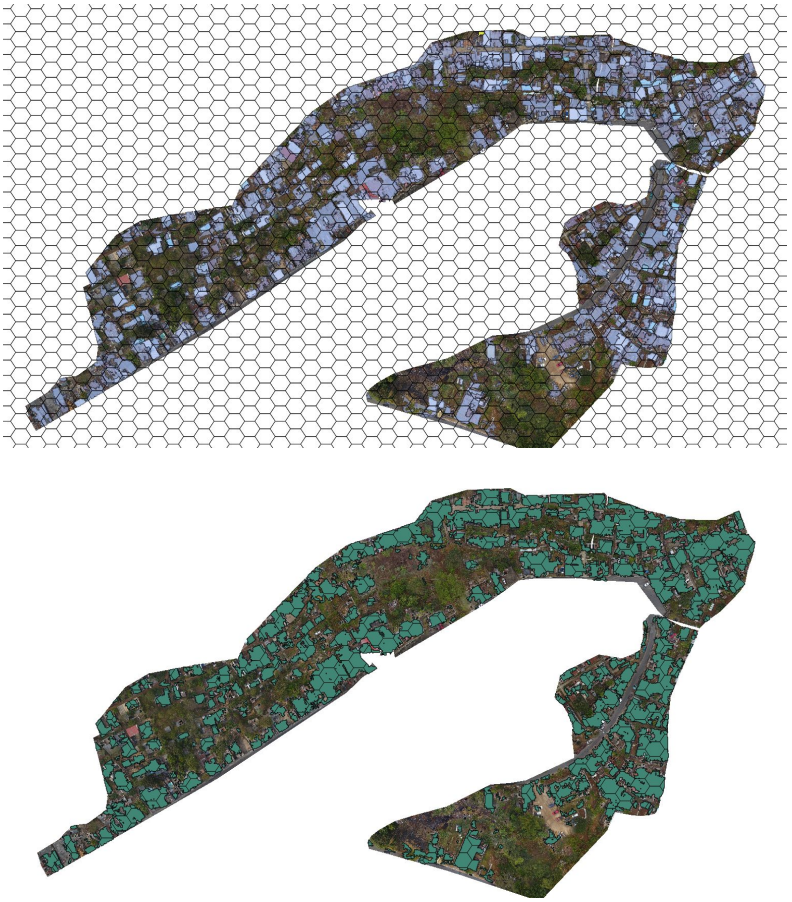


Example: Supervised classification of households in Arzu. Households are classified as blue via QGIS' Semi-automatic Classification plugin.

c) Creation of sampling-frame based on Sample Units (SUs).

The next step consists of the establishment of a sampling frame based on an hexagonal grid. Since, so far, it is not possible to determine N, the number of households, we shall estimate such number via an estimator $\overline{N}$, which in turn will be based on N_{SU} Sample Units and η expected number of households per Sample Unit. In turn, η shall be estimated via $\overline{\eta}$, obtained via field observations.

The grid will be of fixed size which will be expected to contain more than one house. Once the grid has been established, it will be merged with the households layer obtained in the previous step. The hexagons will be both the sampled and surveyed units, i.e., if an hexagon is selected in the sample, the pollster will proceed to survey all the houses that are mostly [*this requires a more precise definition*] contained in the hexagon. As part of the survey itself, the pollster shall write down the number of households which are mainly contained in the hexagon.



Creation of grid and cut-out with classified households.

d) Sampling based on Sampling Units (SUs).

The sampling methodology will use probability proportional to the area covered by a physical structure. The user must calculate the percentage the hexagon's area covered by households before sampling them. Let α_i the area covered by households in hexagon i, and β_i the area covered by hexagon H_i. In random sampling, hexagon H_i would be selected with probability $p_i^{RS} = \frac{\beta_i}{\Sigma\beta_i}$; however, in sampling proportional to expected size, hexagon i will be selected with probability:

$$p_i^{PS} = \left[\frac{\alpha_i}{\beta_i}\right] / \left[\Sigma\frac{\alpha_i}{\beta_i}\right]$$

Notice that N_{SU}, the number of SUs, must not necessarily be equal to the number of hexagons. Given that SUs consist of areas where we expect households to be present, and hexagons are

created regardless of the absence of structures, it must follow that all Hexagons must contain non-negligible household areas. In other words, SU, the set of all SUs H_i, must comply to:

$$SU = \{H_i \mid \alpha_i > 0\}$$

The number of houses to surver, n_{SU} must be calculated according to N_{SU} and the desired survey accuracy.

Example: Hexagons expected to contain households

References

[1] http://www.ub.edu/medame/foro_ptdr/m4/SGONZALEZ.pdf
[2] http://www.techo.org/paises/mexico/opina/los-asentamientos-humanos-irregulares-una-mirada-hacia-su-definicion/
[3] https://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf