Harnessing the predictive power of community workshops, geospatial data, and Bayesian statistics to address census omissions in remote areas of Colombia

Lina Maria Sanchez-Cespedes[1], Glenn Harry Amaya Cruz[1], Mariana Ospina Bohórquez[1], Douglas Ryan Leasure[2,3], Natalia Tejedor-Garavito[2]

[1] National Administrative Department of Statistics (DANE), Bogotá, Colombia
[2] WorldPop, Department of Geography and Environmental Sciences, University of Southampton, UK
[3] Leverhulme Centre for Demographic Science, Department of Sociology, University of Oxford, UK

Theme: Spatial Demography and Human Geography

Keywords (4): Bayesian statistics; population and housing census; GIS and remote sensing; rural community engagement

Abstract (200 words)

A full coverage national population and housing census or estimates of degree of completeness by municipality are essential to identify numbers of people and dwellings that are key for government planning and decision making. In Colombia, this is challenging for remote regions with low population densities, large territorial extents, and insecurity in some areas. Considering the importance of estimating census completeness at the municipality level and the flexibility of hierarchical Bayesian models to estimate census omissions, this study explores differences among three competing models. These combined population data from the 2018 census, social cartography community workshops, with independent variables derived from GIS and remote sensing data to estimate census omissions for remote regions with limited information. As training data, we used census results of nearby fully enumerated areas. We assessed covariate effects, out-of-sample prediction accuracy, and uncertainty intervals using 10-fold cross-validation. The model with the best prediction accuracy combined information from *community-based workshops* and *satellite-based building maps* that were fit to *census-based field observations*. These population estimates and uncertainty intervals support government planning in locations not fully accessible to census enumerators.

Theoretical Focus

Census coverage is the degree of completeness of the observed units, dwellings, and people. The total count of these units is not always possible, given some external factors that impact the statistical operation. Having a degree of completeness by municipality is very important because knowing the number of people is a key input for government entities for planning and decision-making. Nowadays, Bayesian Hierarchical Models (BHMs) are being used to estimate census completeness because of their flexibility to formulate bespoke models to estimate the approximate population of an area.

In Colombia there are areas, mainly in the regions of the Amazonía, Orinoquía and Pacífica, which are characterized by their difficult accessibility, low population density, large territorial extension, and, in addition, some of these have security problems. The sum of these conditions resulted in greater challenges in both the planning and operation of any kind of census. In addition, these areas do not have reliable demographic information, their administrative records are incomplete, and there is no clarity of administrative boundaries between municipalities among their inhabitants. To overcome these challenges, the National Statistics Office of Colombia (DANE) implemented about ninety social cartography workshops (SCW) with ethnic community representatives (indigenous and afro-descendants) of these rural locations, held between 2011 and 2014 and updated in 2016 and 2017.

The objective of the SCW was to collect information on the location and basic characteristics of the ethnic communities or population settlements such as the approximate number of housing units and people. For the 2018 Population and Housing Units Census (2018 PHUC), a data collection method called routes (*rutas* in Spanish) was developed utilizing information from the workshops and other sources (e.g. 3rd National Agricultural Census, territorial planning documents, municipal and departmental development plans). This method of collection consisted of working groups that went through the rural area of some municipalities. The routes were usually along a river and its tributaries, horseshoe paths or logging roads, encompassing an area of influence containing each of the existing communities and settlements.

In addition to the information of the SCW, we use geospatial covariates because of the problems described previously: information deficit and unknown municipality boundaries among inhabitants. However, technology through geospatial data such as those from satellite images provide information, directly or indirectly, of the presence of people in the territory. Thus, by obtaining information about constructions, vegetation, light intensity, etc., we could deduce the presence or absence of people in remote areas. Considering the four aspects discussed previously (importance of census omission, SCW, geospatial covariates and BHMs), the main objective of the study is to show how SCW, geospatial data and BHMs complement one another to obtain reliable estimates of census omission in remote and inaccessible areas. Therefore, results and models obtained by different BHMs are presented to show the importance of considering, on one hand, the knowledge of inhabitants and natives of a region and, on the other, the information from geospatial data.

Data

Social cartography workshops. Ethnic groups occupy approximately 35,215,976 Ha, a third of the national territory. To involve them in the Censuses activities, DANE implemented about ninety SCW with ethnic community representatives (indigenous and afro- descendants) held between 2011 and 2014 for the National Agricultural Census and updated in 2016 and 2017 for the 2018 PHUC. The objective of the SCW was to establish the location of the ethnic communities and their characteristics.Therefore, DANE was able to plan the operational needs in each area (number of census takers and supervisors, costs and times). To achieve it, DANE signed 6 agreements with Afro-Colombian organizations and 8

with indigenous organizations.

For the SCW, the ethnic organization oversaw the logistical aspects and summoned the representatives. They explained the importance of the event and the information that participants had to provide to DANE: number of dwellings, families and people living in each community, access, and costs. During the SCW, the DANE team explained cartography basic concepts and how to geographically locate the communities on the map. Then, assistants were distributed in small groups according to their indigenous reservations, community councils and zones. With clear concepts, each group located their communities on a map. Finally, the groups completed a form by community indicating the number of dwellings, people, etc. Thanks to the SCW, DANE could identify and locate 12,067 communities: 8,010 indigenous, 3,200 afro-colombians and 587 of colonists. This information was used to construct the PHUC frame for routes with the 2014 agricultural census and municipal development plans.

Population Data. We used counts of people and dwellings from the 2018 population and housing census of Colombia, primarily from the Amazonía, Orinoquía, and Pacífica regions. In the census, municipalities from the study regions were divided into operational areas called “routes”, and each of these were divided into “census units”. These were the spatial unit of analysis for our statistical models. In total, there were 394 routes that consisted of 1,302 census units. These routes belonged to 145 municipalities and 23 departments.

Census Coverage. During the census fieldwork, the number of dwellings was verified and controlled by a geographic monitoring system that colored the census units according to the percentage of expected dwellings found. The census units that exceeded more than 90 percent of this indicator were colored in green, those that were in a range from 1 to 90 percent in orange, and the units that were not visited in gray (Fig. 1). We used the green census units (n=508) to train the models, whereas the orange and gray units (n=628 and 166, respectively) are where we would like to predict the population.

Figure 1. Geographic monitoring system. Coverage classification according to the percentage of expected dwellings: green(>90%), orange(1-90%) and grey (not visited). Image source: 2018 Census Geostatistics Division - DANE.

Independent variables. We used geospatial covariates that were available with full coverage across the study area as predictors of population. Based on the information shared by community representatives in the social mapping workshops, an expected number of dwellings and people were assigned to each census unit. We combined these community-based estimates of building counts with satellite-based measurements of total building coverage (i.e. area) from the German Aerospace Center’s World Settlement Footprint 3D. In addition, we derived several covariates from GIS and remote sensing data

sources that included: poverty index, construction index, school density, distance to populated centres, elevation, and intensity of nighttime lights.

Research Methods

We compared three hierarchical Bayesian models using a consistent set of predictor variables across models. All three approaches used the same Poisson model of total population P in location i(i.e. census enumeration area) that incorporated a random intercept to account for correlations among census units from the same department d or municipality m as well as geospatial covariates x_i.

$$P_i \sim Poisson(H_i \rho_i)$$

$$\rho_i \sim LogNormal(\bar{\rho}_i, \sigma)$$

$$\bar{\rho}_i = \alpha_d + \delta_m + \sum_{k=1}^{k} \beta_k x_{k,i}$$

where H_i is the number of buildings (observed for accessible areas during the census) and ρ_i is the average number of people per building. This includes a log-normal regression on ρ_i with a random intercept by department $\alpha_d$and municipalityδ_m, along with our six geospatial covariates. The residual variance term σ quantifies variation in ρ_i (people / building) that was not explained by the predictors. Our six geospatial covariates$x_{k,i}$ were: (1) school density, (2) poverty index, (3) elevation, (4) nighttime lights, (5) distance to populated center, and (6) construction density

Our three models differed in how they estimated the number of buildings H_i. The census recorded counts of buildings and occupied dwellings, although with incomplete coverage in grey and orange areas. We also had community-based estimates of the number of dwellings B_i with full coverage from the social mapping workshops, as well as satellite-based measurements of area covered by buildings C_i from the German Aerospace Center's World Settlement Footprint 3D.

Community model. This model estimated building counts H_i as a function of the community-based estimates $B_i$and the total area $A_i$of each census unit,

$$H_i \sim Poisson(A_i \theta_i)$$

$$\theta_i \sim LogNormal(\bar{\theta}_i, \sigma)$$

$$\bar{\theta}_i = \alpha_d + \delta_m + \gamma_1 log(\frac{B_i}{A_i}) + \sum_{k=1}^{k} \beta_k x_{k,i}$$

where θ_i is building density. All models used the same six geospatial covariates $x_{k,i}$as above.

Satellite model. The number of buildings H_i was estimated as a function of the remotely-sensed building footprint coverage C_i(i.e. total area covered by buildings) within each enumeration area.

$$H_i \sim Poisson(C_i \theta_i)$$

$$\theta_i \sim LogNormal(\bar{\theta}_i, \sigma)$$

$$\bar{\theta}_i = \alpha_d + \delta_m + \gamma_1 log(\frac{C_i}{A_i}) + \gamma_2 log(A_i) + \sum_{k=1}^{k} \beta_k x_{k,i}$$

Unlike the "community model" above, here$\theta_i$is the density of buildings based on the area covered by buildings C_i rather than the total area A_i of the census unit. The satellite-based maps of building coverage constrains the area where dwellings may occur within each census unit significantly.

Combined model. This model combined the community-based estimates of buildings B_i with the satellite-based estimates of building coverage C_i in an attempt to better estimate the observed total building count from the census H_i. This model was the same as the "satellite model", except:

$$\bar{\theta}_i = \alpha_d + \delta_m + \gamma_1 log(\frac{B_i}{C_i}) + \gamma_2 log(\frac{C_i}{A_i}) + \gamma_3 log(A_i) + \sum_{k=1}^{k} \beta_k x_{k,i}$$

It includes community dwelling estimates $B_i$and satellite-based building coverage C_i as predictors.

We used 10-fold out-of-sample cross validation to measure prediction accuracy (*e.g.* r^2) and the robustness of uncertainty intervals. Priors (not shown) were designed to be minimally informative. Models were fit using JAGS software in the R statistical programming environment and convergence was confirmed with standard tests (i.e. $psrf < 1.1$and trace plots).

Expected Findings

Although we expect to see some differences in the effects of predictors on population estimates among models, we hope to identify some predictors that are important across all three models. In particular, we are interested in assessing the predictive power of community-based estimates of dwelling counts versus satellite-based observations of buildings in difficult-to-access remote areas.

We will compare the models based on their predictive performance. We expect the mean posterior predictions from all three models to be relatively unbiased estimators of total population in each coverage area, but we expect the models to achieve different levels of precision. The 95% prediction intervals from all of the models should contain about 95% of the out-of-sample observations from the training data to indicate robust prediction intervals. We will derive probabilistic estimates of total population for census units, routes, municipalities, departments, regions, and the total study area. We will compare the mean predictions and the uncertainty intervals among models and also against other sources of information about the populations in these areas.

This work will provide evidence to help guide decision-making to address census omissions in remote areas of Colombia. We hope to demonstrate the predictive power of communities in these regions that are difficult to access for census enumerations, particularly when combined with GIS and remote sensing data (such as building footprints) that have full coverage nationally. The Bayesian approach that we adopted for this work provides flexibility to develop bespoke models to maximize the inference power from existing data and also to objectively quantify the statistical uncertainty associated with any population estimates. Not only is this important for making informed decisions from imperfect data, but also to build trust in the methods being applied to complement census field work in hard-to-reach rural areas.