


## Research Articles

# Evaluation of residential structures not covered by aerial photographs used to generate a sampling frame – Nueva Santa Rosa, Guatemala

Jeffrey M. Switchenko<sup>1</sup> <sup>a</sup>, Sharon L. Roy<sup>1</sup> <sup>b</sup>, Fredy Muñoz<sup>2</sup>, Gerard Lopez<sup>3</sup>, Jose G. Rivera<sup>2</sup>, Victoria M. Cuéllar<sup>1</sup>, Patricia Juliao<sup>4</sup>, Beatriz López<sup>2</sup>, Andrew Thornton<sup>5</sup>, Jaymin C. Patel<sup>6</sup>, Maricruz Alvarez<sup>2</sup>, Lisette Reyes<sup>7</sup>, Gordana Derado<sup>1</sup>, Wences Arvelo<sup>4</sup>, Kim A. Lindblade<sup>4</sup>

<sup>1</sup> Division of Foodborne, Waterborne, and Environmental Diseases, Centers for Disease Control and Prevention (CDC), Atlanta, Georgia, USA, <sup>2</sup> Centro de Estudios en Salud, Universidad del Valle de Guatemala, Guatemala City, Guatemala, <sup>3</sup> CDC Regional Office for Central America, Guatemala City, Guatemala, <sup>4</sup> CDC Regional Office for Central America, Guatemala City, Guatemala; Centers for Disease Control and Prevention (CDC), Center for Global Health, International Emerging Infections Program, Atlanta, Georgia, USA, <sup>5</sup> Centers for Disease Control and Prevention (CDC), Division of Foodborne, Waterborne, and Environmental Diseases, Atlanta, Georgia, USA; Rollins School of Public Health, Emory University, Atlanta, Georgia, USA, <sup>6</sup> CDC Regional Office for Central America, Guatemala City, Guatemala; Rollins School of Public Health, Emory University, Atlanta, Georgia, USA, <sup>7</sup> Health Area of Santa Rosa, Ministry of Public Health and Social Welfare, Santa Rosa, Guatemala

Keywords: epidemiology, sampling frame, satellite imagery, guatemala, geographic information systems

<https://doi.org/10.29392/001c.24585>

## Journal of Global Health Reports

Vol. 5, 2021

### Background

Aerial images are being used more often to map residential structures on the ground in a study area (the sample frame). However, non-coverage bias associated with overhead imagery has not been fully explored. Non-coverage occurs when residential structures are not included in a particular sampling frame. Our study aimed to evaluate non-coverage bias and sensitivity of an aerial photograph methodology in Nueva Santa Rosa, Guatemala, which was used to generate the sampling frame for a larger cross-sectional survey of sanitation, disease, and water quality.

### Methods

High-resolution aerial photographs of Nueva Santa Rosa were overlaid with a grid, and roof images were geo-located within randomly sampled cells, dichotomized by population as very high-density (VHD) or non-VHD. Roofs found on-site were compared to roofs found in photographs to evaluate the numbers and sizes of residences excluded from the sampling frame. Non-coverage proportions were estimated, and sensitivity and specificity were assessed.

### Results

There was no statistically significant difference (1.2%; 95% confidence interval, CI= -12.1-14.6) in non-coverage proportion between VHD segments (39.6%) and non-VHD cells (38.4%). Roof-size range sensitivity and specificity were 66.4% (95% CI=57.6–74.2) and 69.4% (95% CI=54.4–81.3).

### Conclusions

Approximately one-third of residential roofs were missed, perhaps due to outdated photographs. No substantial bias concerning population density appeared to influence our sampling frame. Further assessment of non-coverage bias, possibly expanding the roof size range to modify sensitivity and specificity, should be performed to generate geographically based best practices for overhead-image use.

Unavailable, unreliable, or incomplete household maps or lists are common challenges to epidemiologic research, particularly in resource-limited settings. The time, effort, and/or money required to develop or acquire such maps or

<sup>a</sup> Joint First Authors

<sup>b</sup> Joint First Authors

lists has often compelled researchers to opt for variants of the random walk,<sup>1,2</sup> but the outputs of such activities are not necessarily true probability-based samples.<sup>3,4</sup> A probability sample must rely on a map or list from which households may be randomly sampled (the sampling frame).<sup>5</sup> Satellite and aerial imagery that are available freely or at a modest price provide a possible alternative to traditional maps and lists. One may employ geospatial mapping of all potential residential-associated (PRA) roofs on satellite or aerial images within a chosen geographic area to develop a map for simple random sampling of households. This method has been used in resource-poor areas with inadequate maps, census data, and infrastructure, or for surveying vulnerable populations in insecure environments.<sup>6–16</sup> It has required advanced mapping and spatial software, such as ArcGIS or AutoCAD, to identify PRA roofs in overhead images obtained using Quickbird, Google Earth, or IKONOS.<sup>6,7,9,12–14</sup> To estimate coverage using overhead imagery, ground teams are often sent with global positioning system (GPS) devices to pinpoint the identified roofs on-site and verify their status as residences or non-residential structures.<sup>7–9,11–16</sup> In two studies for which time frames were provided,<sup>7,13</sup> residential structures made up approximately 95% of the roofs that were successfully located on the ground using overhead images < 3 years old.

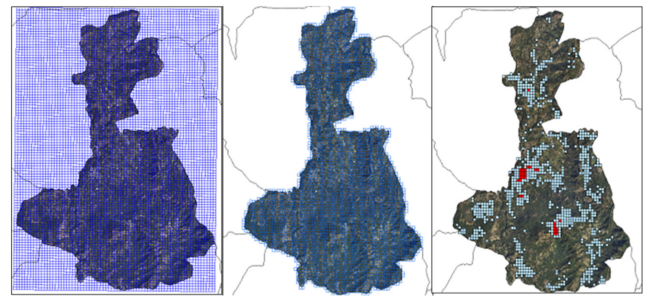
While such success in locating residential structures is encouraging, few studies have estimated the proportion of residential structures not captured in the images.<sup>16</sup> Non-coverage arises when residential structures are not included in a particular sampling frame and have no chance of being sampled. Consequently, their absence may affect the representativeness of the study. When using overhead imagery, non-coverage likely occurs for several reasons and may vary by location if (1) new roofs have been added or existing roofs have been removed since the images were taken, (2) roofs have been obscured in the images (e.g., by foliage), (3) roofs are outside the designated PRA-roof size range or are thought to be non-residential and are not mapped, or (4) the image resolution is inadequate to distinguish multiple roofs (and therefore potentially multiple households) from single roofs (e.g., in high-density urban areas).

The extent of non-coverage bias associated with overhead imagery has not been fully explored. Therefore, we conducted a sub-study in the *municipio* of Nueva Santa Rosa in Guatemala in August–September 2010 to evaluate the non-coverage bias and sensitivity of an aerial photograph methodology used to generate the sampling frame for a larger cross-sectional survey of diarrhea and soil-transmitted helminthiasis prevalence and associated water, sanitation, and hygiene risk factors.<sup>17</sup>

## METHODS

### SITE

Nueva Santa Rosa (NSR) is a *municipio* (municipality) in the *departamento* (state) of Santa Rosa, approximately 45 kilometers southeast of Guatemala City, the capital of Guatemala. NSR is a sparsely populated mountainous area with three main urban areas linked by paved roads; the remainder of NSR is served by dirt roads. In 2010, NSR



**Figure 1. A composite of aerial photographs of the *municipio* of Nueva Santa Rosa (NSR), Guatemala.**

(A) Under a grid of 200m x 200m cells; (B) Cells without contact with NSR removed; (C) Demonstrating cells with population densities >5000 persons/km<sup>2</sup> (red cells) or <5000 persons/km<sup>2</sup> (blue cells) estimated using the number of possible residential-associated roofs per cell.

had an estimated population of 31,044 people<sup>18</sup> and 5,918 households,<sup>19</sup> as determined by the Instituto Nacional de Estadística. Further vital statistic details for Nueva Santa Rosa in 2010 can be found at the following link (<https://www.ine.gob.gt/sistema/uploads/2013/12/10/MVd-hUf5YNLubC3ZikAABJekA0ettQNw1.pdf>).

### MAPS

We obtained high-resolution, georeferenced, geometrically corrected (orthorectified) aerial photographs of NSR from Guatemala's Instituto Geográfico Nacional taken in 2006. The photograph pixel resolution was approximately 0.16m<sup>2</sup>, functioning on a 1:10,000 scale. We formatted all photographs with ArcGIS and overlaid them with the 2002 NSR census tract maps to identify the *municipio* boundaries. We created a grid of 200m x 200m cells (0.04km<sup>2</sup>) to cover NSR, overlaid a topographic map with GPS coordinates (Figure 1, panel A), then discarded all cells that did not touch the *municipio* (Figure 1, panel B).

### IDENTIFYING POTENTIAL RESIDENTIAL-ASSOCIATED ROOFS

We defined PRA roofs as human constructions 16–150m<sup>2</sup> that might be residential, similar to a previous study of household enumeration using overhead imagery, where structures between 9–330m<sup>2</sup> were selected for study assessment.<sup>20</sup> Based on investigator knowledge, we believed this was the most appropriate size range for NSR roofs that could represent residential structures such as entire houses, or buildings used for separate household functions such as kitchens, dining rooms, living rooms, and bedrooms. If part of a roof was hidden, we could still measure and categorize it as a PRA roof if two diagonal corners or three corners were visible. We used the length of the shadow cast by the building compared to other buildings of similar roof size to distinguish multi-story buildings. Once identified, we marked PRA roofs with a red dot on the digitized aerial photographs and gave each dot a GPS coordinate. We excluded the following non-residential roofs (≥16m<sup>2</sup>) by shape or based on previous knowledge: churches, community halls, supermarkets, health centers, schools, gov-

ernment buildings, police stations, factories, warehouses, barns, wineries, and circuses. Digitization of the aerial photographs and manual identification, marking, and geo-location of the PRA roofs took 60 person-days of full-time work.

### SAMPLING

We hypothesized that non-coverage bias varied by location. High-density urban NSR areas were generally organized into blocks or segments of compact structures and posed the greatest challenge in separating one roof from the next in the photographs. To ensure we evaluated these challenging urban areas, we sub-divided NSR by population density and chose to set our cut-off at a very high population density of  $\geq 5,000$  persons/km<sup>2</sup>. With a cell size of 0.04km<sup>2</sup>, an estimated average Santa Rosa household size of 4.8 persons per household,<sup>19</sup> and an estimated average urban roof-to-household ratio of 1.25 based on *a priori* investigator opinions, we predicted that cells with very high-density populations would contain  $\geq 52$  PRA roofs. Using this cut-off, we categorized each cell containing  $\geq 52$  PRA roofs as a very-high-density (VHD) cell and the rest as non-VHD cells; empty cells were excluded from the sampling frame (Figure 1, panel C).

To achieve a fairly even distribution of PRA roofs between the two groups, we selected twice as many non-VHD cells as VHD segments to account for the higher density of roofs in the VHD segments. We randomly selected 10 non-VHD cells from a list of all non-VHD cells. VHD cells underwent a two-step selection process. First, we randomly selected five VHD cells from a list of all VHD cells. Next, we further divided these five VHD cells into blocks or segments, each containing approximately equivalent but  $< 20$  PRA roofs. We used natural divisions like roads or rivers to guide segmentation, where possible, and included the entire area within each VHD cell in the segmentation process (Figure 2). We randomly selected one segment to represent each of the five VHD cells. We could only evaluate 10 non-VHD cells and five VHD segments with the available time and manpower.

In the selected non-VHD cells and VHD segments, we marked all non-PRA roofs that were either too small ( $4\text{m}^2 - < 16\text{m}^2$ ), too large ( $> 150\text{m}^2$ ), or otherwise excluded (e.g., known to be a church) with green dots (Figure 3). No green dots were given to roofs  $< 4\text{m}^2$  under the assumptions that they were too small to be stand-alone houses and therefore their associated households would be represented in the sampling frame by other larger roofs.

### GROUND WORK

Using personal digital assistants (PDAs) and copies of the aerial photographs, we attempted to locate red-dot and green-dot roofs by their GPS coordinates and positions on the photographs. On the ground, we generated GPS coordinates for roofs we found and to which we had access that had no corresponding images on the photographs (i.e., no-dot roofs), either because they were newly built since the photographs were taken or were obscured in the photographs.



**Figure 2. Segments within a very-high-density cell overlaying a 2006 aerial photograph of the *municipio* of Nueva Santa Rosa, Guatemala.**



**Figure 3. Two examples of non-very-high-density cells overlaying a 2006 aerial photograph of the *municipio* of Nueva Santa Rosa, Guatemala.**

Red-dots indicate potential residential-associated (PRA) roofs  $16\text{m}^2$  to  $150\text{m}^2$  in size. Green dots indicate non-PRA roofs, either too small ( $< 16\text{m}^2$ ), too large ( $> 150\text{m}^2$ ), or otherwise excluded because they were known to investigators to be non-residential structures (e.g., churches, community halls, supermarkets, health centers, schools, government buildings, police stations, factories, warehouses, barns, wineries, or circuses).

### STATISTICAL ANALYSES

Non-coverage was defined as one minus coverage, where coverage was the percent of residential roofs found during the ground work that were red-dot PRA roofs and therefore in the sampling frame. We compared residential proportions between non-VHD cells and VHD segments for green-dot roofs and no-dot roofs using Chi-square tests or Fisher's exact tests, where appropriate. Not all red-dot PRA roofs were visited due to resource constraints so, to estimate non-coverages and confidence intervals, we imputed the residential status (yes or no) of the non-visited red-dot PRA roofs using the Bernoulli distribution where the probability

of imputation as a residential-associated roof vs. non-residential-associated roof was based on the estimated residential rate where the non-visited roof was located (non-VHD cell or VHD segment), i.e., a weighted coin-flip approach. We then resampled with replacement the set of residences, both observed and imputed, and calculated the non-coverage proportion for that sample. This process of imputing and resampling was repeated 10,000 times, to obtain an average non-coverage proportion and 95% empirical confidence interval based on the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles of the calculated non-coverage probabilities, thus capturing both sampling variability and variability based on imputation within our confidence interval estimates. We compared non-coverage probabilities of non-VHD cells and VHD segments using the difference in resampled proportions, and estimated a 95% confidence interval using the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles of the resampled distribution of the difference in non-coverage probabilities. We set statistical significance at 0.05 and performed the analysis using SAS 9.4 (SAS Institute Inc., Cary, NC) and R 4.0.5.<sup>21</sup>

We repeated these analyses using a different population density definition to re-classify cells from a very high density of  $\geq 5000$  persons/km<sup>2</sup> to  $\geq 1000$  persons/km<sup>2</sup>, a figure some have used to define urban areas.<sup>22</sup> This changed the proportion of roofs in high versus low-density cells to allow us to further evaluate the effect of population density on non-coverage.

Finally, we evaluated the sensitivity and specificity of the PRA-roof 16–150m<sup>2</sup> size range used to generate the sampling frame. Because of the 4-year interval between the photographs and the fieldwork, we excluded no-dot roofs (i.e. roofs discovered in the field that were not identified in the photographs) because there was likely a lot of new construction in the 4-year interval. An unbiased analyst retrospectively categorized green-dot roofs into “too large” ( $>150\text{m}^2$ ), “too small” ( $4\text{m}^2\text{--}<16\text{m}^2$ ), or “size unknown” based on visual size estimates in comparison to red-dot PRA roofs sizes using the same aerial photographs used by staff who generated the original maps. The sensitivity of the residential size range used for classification was calculated as the percent of residential roofs within the size range, and the specificity was calculated as the percent of non-residential roofs outside of the size range; confidence intervals were calculated using the efficient-score method corrected for continuity.<sup>23</sup>

## RESULTS

### NON-COVERAGE ANALYSIS

A total of 3,848 cells covered NSR across 144.3 km<sup>2</sup>. We identified 10,770 red-dot PRA roofs in NSR. The non-coverage sub-study included 193 PRA roofs: 102 in the 10 non-VHD cells and 91 in the five VHD segments (Table 1). One randomly selected non-VHD cell in a very remote mountainous area could not physically be reached; we replaced it with the next non-VHD cell in the sampling frame list.

We looked for a convenience sample of 122 (63.2%) of 193 red-dot PRA roofs and found 102 (83.6%), of which 87 (85.3%) were residentially associated; we did not look for 71 (36.8%) PRA roofs due to personnel and time constraints. Of

36 PRA roofs not visited in non-VHD cells, 29 (80.6%) were from two cells; similarly of 35 PRA roofs not visited in VHD segments, 31 (88.6%) were from two segments. The structure types for residential and non-residential buildings are listed in Tables 2 and 3, respectively.

We also mapped 102 green-dot roofs either  $4\text{m}^2\text{--}<16\text{m}^2$  or  $>150\text{m}^2$ . We searched on the ground for all green-dot roofs and found 81 (79.4%). We found 159 no-dot roofs (Table 1). Of the 240 non-PRA roofs located in the field (81 green-dot roofs + 159 no-dot roofs), we determined 94 (39.2%) to be residential structures, which served a variety of functions and varied in sizes (Tables 4 and 5). Seven of these belonged to households that already had red-dot PRA roofs included in the sampling frame, based on homeowner confirmation. Of the remaining 87 roofs, 53 were in non-VHD cells and 34 were in VHD segments. These roofs were not known to be associated with households that had PRA roofs already in the sampling frame, and we assumed all 87 roofs belonged to separate households. Of these 87 missed residential structures, 30 (34.5%) were within the 16–150m<sup>2</sup> range, 37 (42.5%) were  $<16\text{m}^2$ , 11 (12.6%) were  $>150\text{m}^2$ , and 9 (10.3%) were of unknown size. Correcting for the seven green-dot and no-dot roofs already associated with red-dot PRA roofs, 57.1% (32/56) of green-dot roofs were residential structures in non-VHD cells compared to 54.5% (12/22) of green-dot roofs in VHD segments ( $P=0.84$ ); 23.6% (21/89) of no-dot roofs were residential structures in non-VHD cells, compared to 33.3% (22/66) of no-dot roofs in VHD segments ( $P=0.18$ ).

We determined 87 red-dot PRA roofs to be residential, 63.2% (55/87) in non-VHD cells and 36.8% (32/87) in VHD segments, and found a total of 87 missed green-dot and no-dot residential structures, 60.9% in non-VHD cells and 39.1% in VHD segments. In addition, 71 red-dot PRA roofs were not visited so their residential status could not be determined. Using imputation and resampling techniques, we estimated the coverage proportion for non-VHD cells to be 61.6%, and the non-coverage proportion to be 38.4% (95% confidence interval, CI=30.4–46.8). The coverage proportion for VHD segments was 60.4%, and the non-coverage proportion was 39.6% (95% CI=29.1–50.6). The difference in non-coverage proportion between non-VHD cells and VHD segments was not statistically significant (Difference = 1.2%, 95% CI= -12.1–14.6).

### POPULATION DENSITY SENSITIVITY ANALYSIS

To evaluate the effect of population density, we used an alternative high-density definition of  $\geq 1000$  persons/km<sup>2</sup>; five non-VHD cells were now considered high density. Of the 87 red-dot residential roofs, 82.8% (72/87) were in high-density cells and 17.2% (15/87) were in low-density cells. Of the 87 green-dot/no-dot residential roofs, 90.8% (79/87) were in high-density cells and 9.2% (8/87) were in low-density cells. All 71 PRA roofs that we did not visit were located in high-density cells. The estimated non-coverage proportion in high-density cells was 39.4% (95% CI=32.4–46.4), and the estimated non-coverage proportion in low-density cells was 34.8% (95% CI=17.4–52.2). The difference in non-coverage proportion between high- and low-density cells was not statistically significant (Difference = 4.8%, 95% CI=

**Table 1. Comparison of roofs in the 2006 aerial photographs to those observed in the field, by size and number – Nueva Santa Rosa, Guatemala, 2010**

Roof type	Roof size	Location	Number of roofs identified in photos	Number of roofs looked for in field	Number of roofs found in field	Number of roofs residentially associated
Red-dot (PRA*) roofs	16m <sup>2</sup> –150m <sup>2</sup>	Total	193	122	102	87
		Non-VHD cells	102	66	62	55
		VHD segments	91	56	40	32
Green-dot roofs	4m <sup>2</sup> – <16m <sup>2</sup>	Total	73	73	55	29
		Non-VHD cells	52	52	40	22
		VHD segments	21	21	15	7
	>150m <sup>2</sup>	Total	14	14	13	9
		Non-VHD cells	12	12	11	8
		VHD segments	2	2	2	1
	Size unknown	Total	15	15	13	9
		Non-VHD cells	9	9	8	5
		VHD segments	6	6	5	4
	<b>Total</b>	<b>Total</b>	<b>102</b>	<b>102</b>	<b>81</b>	<b>47</b>
		<b>Non-VHD cells</b>	<b>73</b>	<b>73</b>	<b>59</b>	<b>35</b>
		<b>VHD segments</b>	<b>29</b>	<b>29</b>	<b>22</b>	<b>12</b>
No-dot roofs	<4m <sup>2</sup>	Total	-	-	23	1
		Non-VHD cells	-	-	16	1
		VHD segments	-	-	7	0
	4m <sup>2</sup> –<16m <sup>2</sup>	Total	-	-	62	12
		Non-VHD cells	-	-	43	8
		VHD segments	-	-	19	4
	16m <sup>2</sup> –150m <sup>2</sup>	Total	-	-	68	32
		Non-VHD cells	-	-	30	14
		VHD segments	-	-	38	18
	>150m <sup>2</sup>	Total	-	-	4	2
		Non-VHD cells	-	-	2	1

		VHD segments	-	-	2	1
	Size unknown	Total	-	-	2	0
		Non-VHD cells	-	-	1	0
		VHD segments	-	-	1	0
	Total	Total	-	-	159	47
		Non-VHD cells	-	-	92	24
		VHD segments	-	-	67	23

\* PRA – Potential residential-associated roof; VHD – very high density.

**Table 2. Structure types for red-dot (PRA\*) roofs, among residential buildings – Nueva Santa Rosa, Guatemala, 2010.**

Structure type	Frequency
<b>House</b>	<b>61</b>
House	52
House/Store	4
House/Room	2
House/Apartment	1
House/Bedroom	1
House/Room	1
<b>Bedroom</b>	<b>12</b>
Bedroom(s)	11
Bedroom/Kitchen	1
<b>Room</b>	<b>7</b>
Room	6
Room/Storage area	1
<b>Kitchen</b>	<b>5</b>
Kitchen	4
Kitchen/Dining Room	1
<b>Other</b>	<b>1</b>
Patio	1
<b>Total</b>	<b>86<sup>†</sup></b>

\* Potential residential-associated roof

† One non-residential structure is missing a description

-16.6–24.8).

#### ROOF SIZE RANGE SENSITIVITY AND SPECIFICITY

We determined that 47 of 102 green-dot roofs either 4m<sup>2</sup>–<16m<sup>2</sup> or >150 m<sup>2</sup> were residential roofs, although three belonged to households that already had red-dot PRA roofs included in the sampling frame. Comparing these green-dot roofs to the 122 red-dot PRA roofs we looked for, 87 of which were residential, we calculated the sensitivity

of the 16–150m<sup>2</sup> size range used for red-dot PRA roofs to be 66.4% (95% CI=57.6–74.2) and the specificity to be 69.4% (95% CI=54.4–81.3).

#### DISCUSSION

Using overhead images to build a sampling frame has several advantages. It permits simple random sampling of roofs and associated households, improving study power

**Table 3. Structure types for red-dot (PRA\*) roofs, among non-residential buildings – Nueva Santa Rosa, Guatemala, 2010.**

Structure Type	Frequency
Storage area	6
School	2
Store	2
Chicken coop/storage area	1
Church	1
Mill	1
Room	1
<b>Total</b>	<b>14<sup>†</sup></b>

\* Potential residential-associated roof

† One non-residential structure is missing a description

**Table 4. Types and sizes of residential structures excluded from the sampling frame (green-dot roofs plus no-dot roofs) generated using 2006 aerial photographs – Nueva Santa Rosa, Guatemala, 2010.**

Structure type	Size of non-PRA* roofs excluded from sampling frame				
	<4m <sup>2</sup>	4m <sup>2</sup> –<16m <sup>2</sup>	16m <sup>2</sup> –150m <sup>2</sup>	>150m <sup>2</sup>	Size unknown
House	0	5	20	11	4
Kitchen	0	19	7	0	4
Rooms	0	5	1	0	1
Bedrooms	0	10	4	0	0
Other	1	1	0	0	0
Missing	0	1	0	0	0
<b>Total</b>	<b>1</b>	<b>41</b>	<b>32</b>	<b>11</b>	<b>9</b>

\* PRA roof = potential residential-associated (red-dot) roof. Non-PRA roofs excluded from the sampling frame include those roofs found on the ground that were not seen in the aerial photographs (no-dot roofs) or those identified in the photographs that were either too small (<16m<sup>2</sup>), too large (>150m<sup>2</sup>), or otherwise excluded because they were known to investigators to be non-residential structures (e.g., churches, community halls, supermarkets, health centers, schools, government buildings, police stations, factories, warehouses, barns, wineries, or circuses) (green-dot roofs). The size of the roof was considered unknown if we were unable to determine whether it was “too small” or “too large” based on the green dots in the aerial photographs, or if we were unable to compute its square area if it was a no-dot roof and its roof dimensions were not available.

and reducing reliance on cluster sampling. Consequently, this methodology could eliminate the subjectivity and convenience bias associated with selecting households in the field. It also creates a sampling frame for subsequent research and surveillance activities and provides geographic coordinates for the exploration of spatial relationships. With increasing use of overhead imagery, one must consider *how much* is missed with the usage of that technology.

Our study also revealed disadvantages of overhead imagery, some of which may not be apparent from the way overhead imagery is presented and analyzed in the literature. Although we were able to find a PRA roof on the ground associated with 102 out of 122 selected PRA roofs (84%), only 87 (71%) of these roofs were associated with residences. This proportion of roofs associated with residential structures was lower than the 94%–97% found in other studies.<sup>7,9,11</sup> Some of this discrepancy was anticipated given the age of the photographs we used (4 years in our study vs. < 3 years in others), which was a significant limitation of the study. However, it has been shown previously that images of that age (>4 years) may not introduce

significant geographic bias when constructing a random sampling of households.<sup>24</sup> This provides further justification for utilizing aerial photographs despite a gap between the time in which the images were captured and a ground study was conducted.

The high rates of identification reported in the literature may be over-estimates because they discount the residential structures not seen on the images and, therefore, not searched for in the field. Through our non-coverage assessment, we identified 87 non-PRA (green-dot/no-dot) residential roofs that were missed because of outdated photographs, new construction, and the specified PRA-roof size range. The sensitivity was 66.4% and specificity 69.4%. Therefore, future studies in Guatemala using overhead imagery might consider expanding both ends of the PRA-roof size range. Population density or urbanization did not seem to play a significant role in non-coverage error, as there were no statistical differences in the non-coverage error between non-VHD cells and VHD segments or between high- and low-density cells.

Our sampling frame was restricted by limits in photo-

graph resolution, which could have resulted in bias against selecting smaller roofs. Failure to identify small roofs may have had a cascade of effects, such as potentially introducing a socioeconomic bias or reducing the chances of household selection (which was proportional to the number of roofs in the sampling frame). However, only 12% (5/41) of roofs <16m<sup>2</sup> covered small houses versus stand-alone rooms (e.g., kitchens, bedrooms) (Table 4), suggesting that the resolution of the photographs may not have substantially biased against poor families who might only have afforded one dwelling with a single small roof versus more affluent families who could afford to have separate buildings for different household functions (e.g., one building for the kitchen, a separate building for the bedrooms, etc.).

Resource constraints restricted our time and choices resulting in several study limitations. As noted, the age of the photographs at the time of ground work (4 years) was a significant limitation. In addition, although the cells and segments were selected at random, the numbers of each were set based on limited resources and an unmeasured assumption that roof density would be twice as great in urban versus rural areas. This may have introduced bias towards urban or rural, although no significant difference was observed. We also imputed data for 71 PRA roofs we could not search for on the ground. There were 94 green-dot and no-dot residential roofs found on the ground; seven were associated with PRA roofs already in the sampling frame but there were no resources or time left to interview the owners of the other 87 structures. This information would potentially have modified the numbers used in our calculations and decreased non-coverage and changed sensitivity and specificity values. Nevertheless, quantifiable data were generated from this sub-study from which other researchers can draw conclusions and lessons when designing their own sampling frames using overhead imagery.

Finally, time and cost were important considerations. Geocoding of roofs has the advantage that field staff is not required to be present at the study site to perform the initial household identification and sampling, although overhead imagery is needed. The GPS coordinates can be used by interviewers to navigate to the correct destinations, and this reduces the bias towards selecting and visiting households in the field with good access to roads. Although it took 60 person-days of full-time work to digitize photographs and geocode red-dot roofs, it was possible to initiate this process five months before any interviewer went to the field. As the photograph digitization was a limitation, it should be noted that more advanced methods for object detection and pixel classification can now be handled via geospatial deep learning within ArcGIS<sup>25</sup>. Artificial intelligence (AI) could be applied to reduce the amount of time spent geocoding roofs, thereby enhancing our approach for estimating non-coverage. In addition, although aerial photography was utilized for this study, future researchers could consider quantifying non-coverage with satellite imagery. Households for the larger cross-sectional survey were selected before the pilot study began, which greatly facilitated logistical planning for the study. It reduced costs compared to sampling in the field, which would have required our staff of 20 interviewers, supervisors, drivers, and vehicles to extend their time in the field because of the time

**Table 5. Structure types for non-potential residential-associated (non-PRA\*) roofs (green-dot roofs plus no-dot roofs), among residential buildings – Nueva Santa Rosa, Guatemala, 2010.**

Structure Type	Frequency
<b>House</b>	<b>40</b>
House	38
House/Kitchen	2
<b>Kitchen</b>	<b>30<sup>†</sup></b>
Kitchen	27
Kitchen/Bedroom	2
Kitchen/Room	1
<b>Bedroom</b>	<b>16<sup>†</sup></b>
Bedroom	9
Bedroom/Bathroom	2
Bedroom/Kitchen	2
Bedrooms (apartments)	1
Bedrooms/Storage area	1
Half a bedroom	1
<b>Room</b>	<b>7</b>
Room	6
Room/Bathrooms	1
<b>Other</b>	<b>2</b>
Entrance to a house	1
Roof over an oven	1
<b>Total</b>	<b>93</b>

\* PRA roof = potential residential-associated (red-dot) roof. Non-PRA roofs excluded from the sampling frame include those roofs found on the ground that were not seen in the aerial photographs (no-dot roofs) or those identified in the photographs that were either too small (<16m<sup>2</sup>), too large (>150m<sup>2</sup>), or otherwise excluded because they were known to investigators to be non-residential structures (e.g., churches, community halls, supermarkets, health centers, schools, government buildings, police stations, factories, warehouses, barns, wineries, or circuses) (green-dot roofs).

<sup>†</sup> Two roofs qualified as both a bedroom and a kitchen (bedroom/kitchen). As a result, two roofs were subtracted from the total, although individually both roofs are included in the bedroom count and kitchen count.

needed for this activity and the time needed to travel across NSR. This study and the usage of GIS highlights the need to incorporate this technology to enhance public health approaches in low income countries.

## CONCLUSIONS

To our knowledge, this is the first study providing data to assess non-coverage bias associated with the use of overhead images to develop sampling frames for public health research. The overall non-coverage proportion for a sub-study evaluating the use of 4-year-old aerial photographs to generate a simple random sample of PRA roofs was 38.4% for non-VHD cells and 39.6% for VHD segments. The sensitivity for the PRA-roof size definition of 16–150m<sup>2</sup> was 66.4% with a specificity of 69.4%. Although these values are less than ideal, we conclude that there appears to be no sub-

stantial bias in coverage with regards to population density or socioeconomic status. Therefore, we believe we can proceed with analyzing other study results based on data relying on random sampling from this sampling frame without concern about significant systematic sampling bias.

We have presented just one of what we believe are multiple protocols whereby overhead imagery can be used to develop a multi-use sampling frame in resource-poor regions. This study is not a definitive evaluation of one such method but rather we hope it will stimulate further assessments of non-coverage bias in a variety of locations and situations to make these increasingly popular aerial-imagery methodologies more statistically robust and generate geographically based recommendations and best practices for identifying residential structures on overhead images.

#### ACKNOWLEDGEMENTS

We thank the many people involved in organizing and carrying out this study, including the Guatemala Ministry of Public Health and Social Welfare in Guatemala City and Nueva Santa Rosa; the leadership in the city and towns of Nueva Santa Rosa; Allen Hightower formerly of the Centers for Disease Control and Prevention; and most especially the gracious families in Nueva Santa Rosa who opened their homes for this survey and whose participation made this evaluation possible. We would also like to thank Kelly Squires for her assistance with secondary data analyses.

#### ETHICS APPROVAL

The protocol for this sub-study was embedded within the protocol for a larger cross-sectional study and this larger protocol was submitted for ethical review. The protocol was approved by the Ethics Committee of the Universidad del Valle de Guatemala (UVG) in Guatemala City, Guatemala (protocol 038-04-2010, approval date July 19, 2010) and the Institutional Review Board of the Centers for Disease Control and Prevention (CDC) in Atlanta GA, USA (protocol 5936, approval date June 18, 2010).

#### FUNDING

This project was a collaboration between the Centro de Estudios en Salud, Universidad del Valle de Guatemala, the Centers for Disease Control and Prevention Regional Office for Central America and Panama, specifically the International Emerging Infections Program, and the Guatemala Ministry of Public Health and Social Welfare. This study was funded by the Centers for Disease Control and Prevention (CDC) Global Disease Detection Program.

#### AUTHORSHIP CONTRIBUTIONS

KL, PJ, and SR conceived of the study; AT, BL, FM, GL, KL, LR, MA, PJ, SR, VC, and WA designed the study protocol; FM, GL, and JR developed the aerial image maps and sampling frames and managed the data; AT, BL, FM, GL, JP, KL, LR, MA, PJ, and VC carried out the field work; JS analyzed and interpreted the data for this sub-study with input from AT, FM, GD, SR, and VC; JS and SR drafted this manuscript with input from AT, FM, GD, and VC; all authors read and approved the final manuscript. GD, JS, and SR are the guarantors of the paper.

#### COMPETING INTERESTS

The authors completed the Unified Competing Interest form at [www.icmje.org/coi\\_disclosure.pdf](http://www.icmje.org/coi_disclosure.pdf) (available upon request from the corresponding author), and declare no conflicts of interest.

#### CORRESPONDENCE TO:

Jeffrey M. Switchenko, Ph.D.  
Address: 1518 Clifton Road NE, Atlanta GA, 30322;  
Tel: 404-778-4157;  
Email: [jswitch@emory.edu](mailto:jswitch@emory.edu)

Submitted: March 08, 2021 GMT, Accepted: May 21, 2021 GMT



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

## REFERENCES

1. World Health Organization. Training for Mid-Level Managers (MLM)—Module 7: The EPI Coverage Survey. Accessed June 10, 2016. [http://apps.who.int/iris/bitstream/10665/70184/7/WHO\\_IVB\\_08.07\\_eng.pdf](http://apps.who.int/iris/bitstream/10665/70184/7/WHO_IVB_08.07_eng.pdf)
2. Henderson RH, Sundaresan T. Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling methods. *Bull World Health Organ.* 1982;60:253-260.
3. Turner AG, Magnani RJ, Shuaib M. A not quite as quick but much cleaner alternative to the Expanded Programme on Immunization (EPI) Cluster Survey design. *Int J Epidemiol.* 1996;25:198-203. doi:10.1093/ije/25.1.198
4. Bennett S, Woods T, Liyanage WM, Smith D. A simplified general method for cluster-sample surveys of health in developing countries. *World Health Stat Q.* 1991;44:98-106.
5. World Health Organization. *Immunization Coverage Cluster Survey: Reference Manual*. Department of Immunization, Vaccine and Biologicals, World Health Organization; 2005.
6. Lowther SA, Curriero FC, Kalish BT, Shields TM, Monze M, Moss WJ. Population immunity to measles virus and the effect of HIV-1 infection after a mass measles vaccination campaign in Lusaka, Zambia: a cross-sectional survey. *Lancet.* 2009;373:1025-1032. doi:10.1016/s0140-6736(09)60142-2
7. Lowther SA, Curriero FC, Shields T, Ahmed S, Monze M, Moss WJ. Feasibility of satellite image-based sampling for a health survey among urban township of Lusaka, Zambia. *Trop Med Int Health.* 2009;14:70-78. doi:10.1111/j.1365-3156.2008.02185.x
8. Moss WJ, Hamapumbu H, Kobayashi T, et al. Use of remote sensing to identify spatial risk factors for malaria in a region of declining transmission: a cross-sectional and longitudinal community study. *Malar J.* 2011;10:163. doi:10.1186/1475-2875-10-163
9. Pearson AL, Rzotkiewicz A, Zwickle A. Using remote, spatial techniques to select a random household sample in a dispersed, semi-nomadic pastoral community: utility for a longitudinal health and demographic surveillance system. *Int J Health Geogr.* 2015;14:33. doi:10.1186/s12942-015-0026-4
10. Sutcliffe CG, Kobayashi T, Hamapumbu H, et al. Changing individual-level risk factors for malaria with declining transmission in southern Zambia: a cross-sectional study. *Malar J.* 2011;10:324. doi:10.1186/1475-2875-10-324
11. Luquero FJ, Banga CN, Remartinez D, Palma PP, Baron E, Grais RF. Cholera epidemic in Guinea-Bissau (2008): The importance of “place.” *PLoS One.* 2011;6:e19005.
12. Ali M, Rasool S, Park J. Use of satellite imagery in constructing a household GIS database for health studies in Karachi, Pakistan. *Int J Health Geogr.* 2004;3:20.
13. Escamilla V, Emch M, Dandalo L, Miller WC, Martinson F, Hoffman I. Sampling at community level by using satellite imagery and geographical analysis. *Bull World Health Organ.* 2014;92(9):690-694. doi:10.2471/blt.14.140756
14. Lin Y, Kuwayama DP. Using satellite imagery and GPS technology to create random sampling frames in high risk environments. *International Journal of Surgery.* 2016;32:123-128. doi:10.1016/j.ijsu.2016.06.044
15. Wagenaar BH, Augusto O, Ásbjörnsdóttir K, et al. Developing a representative community health survey sampling frame using open-source remote satellite imagery in Mozambique. *Int J Health Geogr.* 2018;17(37):1-13. doi:10.1186/s12942-018-0158-4
16. Bridges DJ, Pollard D, Winters AM, et al. Accuracy and impact of spatial aids based upon satellite enumeration to improve indoor residual spraying spatial coverage. *Malar J.* 2018;17(93):1-8. doi:10.1186/s12936-018-2236-2
17. Matanock A, Lu X, Derado G, et al. Association of water quality with soil-transmitted helminthiasis and diarrhea in Nueva Santa Rosa, Guatemala, 2010. *Journal of Health and Water.* 2018;16(5):724-736. doi:10.2166/wh.2018.207
18. Instituto Nacional de Estadística. *Boletín Informativo Departamento de Santa Rosa Volumen.* 2010;4(4). Accessed June 13, 2016. <http://www.inec.gov.gt/sistema/uploads/2013/12/10/MVdhUf5YNLubC3ZikAABJekA0ettQNw1.pdf>
19. Instituto Nacional de Estadística. Censos Nacionales XI de Población y VI de Habitación 2002—Características de la Población y de los Locales de Habitación Censados. Accessed June 13, 2016. <https://www.inec.gov.gt/sistema/uploads/2014/02/20/jZqeGe1H9WdUDngYXkWt3GIhUUQCukcg.pdf>

20. Kamanga A, Renn S, Pollard D, et al. Open-source satellite enumeration to map households: planning and targeting indoor residual spraying for malaria. *Malar J*. 2015;14(345):1-7. [doi:10.1186/s12936-015-0831-z](https://doi.org/10.1186/s12936-015-0831-z)
21. R: *A Language and Environment for Statistical Computing, Reference Index Version 4.0.5*. R Core Development Team; 2021.
22. OECD. *Redefining “Urban”: A New Way to Measure Metropolitan Areas.*; 2012. [doi:10.1787/9789264174108-en](https://doi.org/10.1787/9789264174108-en)
23. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med*. 1998;17:857-872. [doi:10.1002/\(sici\)1097-0258\(19980430\)17:8](https://doi.org/10.1002/(sici)1097-0258(19980430)17:8)
24. Shields T, Pinchoff J, Lubinda J, Hamapumbu H, Searle K, Kobayashi T, et al. Spatial and temporal changes in household structure locations using high-resolution satellite imagery for population assessment: an analysis in southern Zambia, 2006-11. *Geospatial Health*. 2016;11(140):144-150.
25. ArcGIS API for Python. Geospatial deep learning with arcgis.learn. Accessed May 14, 2021. <https://developers.arcgis.com/python/guide/geospatial-deep-learning/>