Policy Research Working Paper 10040

Data Triangulation Strategies to Design a Representative Household Survey of Hosts and Rohingya Displaced in Cox's Bazar, Bangladesh

Joaquin Endara
Maria Eugenia Genoni
Afsana I. Khan
Walker Kosmidou-Bradley
Juan Muñoz
Nethra Palaniswamy
Tara Vishwanath

WORLD BANK GROUP
Poverty and Equity Global Practice
May 2022

Policy Research Working Paper 10040

Abstract

Obtaining representative information on hosts and displaced populations in a single survey is not straightforward. This paper demonstrates the value of combining traditional and nontraditional sampling frames, geospatial information, and listing exercises to design a representative survey of hosts and Rohingya displaced populations in Cox's Bazar, Bangladesh. The paper applies innovative segmentation techniques using geospatial data to delimit enumeration areas in the absence of updated cartography. The paper also highlights the importance of listing exercises to inform stratification decisions and update population counts.

This paper is a product of the Poverty and Equity Global Practice. 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://www.worldbank.org/prwp. The authors may be contacted at mgenoni@worldbank.org.

Data Triangulation Strategies to Design a Representative Household Survey of Hosts and Rohingya Displaced in Cox's Bazar, Bangladesh[1]

Joaquin Endara

Maria Eugenia Genoni

Afsana I. Khan

Walker Kosmidou-Bradley

Juan Muñoz

Nethra Palaniswamy

Tara Vishwanath

Keywords: Forced displacement, Rohingya, sampling methods, host populations, surveys

JEL codes: C83, I3, F22, R23

[1] Authors are members of the World Bank Poverty and Equity Global Practice. This analysis was undertaken for the design of the Cox's Bazar Panel Survey (CBPS). The CBPS is the result of a partnership between the Yale Macmillan Center Program on Refugees, Forced Displacement, and Humanitarian Responses (Yale Macmillan PRFDHR), the Gender & Adolescence: Global Evidence (GAGE) program, the Poverty and Equity Global Practice of the World Bank. Funding provided by the State and Peacebuilding Fund (SPF). The SPF is a global fund to finance critical development operations and analysis in situations of fragility, conflict, and violence. The SPF is kindly supported by: Australia, Denmark, Germany, The Netherlands, Norway, Sweden, Switzerland, The United Kingdom, as well as IBRD.

1. Introduction

Understanding the implications of large-scale forced displacement events such as refugee influxes or population movements due to natural disasters requires representative and comparable information about the demographic and socio-economic characteristics of the displaced and host populations. This is particularly important when trying to shed light on aspects related to the selectivity of those who are displaced compared to host communities, or the impacts of displacement. Representativeness is central to minimize drawing biased conclusions as decisions to relocate and where to locate in host areas are likely not random. In addition, comparability is key as contrasts between displaced and hosts are a core element of these types of analysis. National household surveys typically exclude displaced populations, particularly at the onset of these events. Humanitarian data collects information on those displaced but these are not designed to answer questions related to selectivity or impacts as they are developed to inform the humanitarian response. One drawback of these data is that they typically exclude representative information on the host populations.

Using the same representative survey to cover both hosts and displaced arises as a desirable strategy, however it is not a straightforward effort. One key challenge is having complete or updated sampling frames to draw the samples of both hosts and arriving migrants. Sampling frames of host populations are typically available but, in many cases, exclude forcibly displaced populations. This is certainly the case when the focus is collecting information right after a mass displacement event as it is unlikely that the sampling frame for that area will be updated to reflect the most recent population. On the other hand, population counts on the recently displaced usually collected by humanitarian agencies may face challenges in providing complete or accurate counts, particularly when there is significant movement of populations within the affected area or when the displaced spread across geographic areas. Additional challenges arise in identifying and counting the displaced in cases where they may not be willing to self-identify.

This paper describes a sampling approach that triangulates different sources of data to design a representative survey in the context of the recent Rohingya displacement from Myanmar to the district of Cox's Bazar in Bangladesh. The Cox's Bazar Panel Survey (CBPS) aims to produce representative information of both the Bangladeshi host population living in the district and the Rohingya displaced living in the district. This paper highlights the challenges and solutions in conducting the CBPS sampling strategy. The approach included combining information from the Bangladesh housing and population census, humanitarian data on Rohingya displaced living in camps, geospatial data and listing exercises. The next section explains the context and the core objectives of the survey. Section 3 describes the sampling exercise. Section 4 summarizes the lessons learned and concludes.

2. Context and objective of the survey

In late August 2017, a large number of Rohingya displaced arrived in the Cox's Bazar district of Bangladesh, fleeing violence from Myanmar. Within a period of four months, some 724,000[2] newly arrived persons joined other Rohingya who had fled earlier waves of violence. By the end of 2018, nearly 2,000 campsites in the Cox's Bazar sub-districts of Teknaf and Ukhia hosted around 912,000 Rohingya.

Compared to other recent forced displaced movements, the 2017 Rohingya influx was unique in its scale and speed. More than 80 percent of all displaced Rohingya who are now in Cox's Bazar arrived within a four-month period.[3] The 2017 wave was also unprecedented in its size relative to the host population. The arriving Rohingya increased the total population of the Cox's Bazar district by 32 percent and the population of Ukhia and Teknaf sub-districts by 154 percent.[4] Today, Rohingya displaced account for as many as 4 out of 5 inhabitants in Ukhia and 4 out of 10 in Teknaf.[5] The concentration of Rohingya displaced in Cox's Bazar is the densest known in the world, reaching 11.7 square meters per person.[6]

This significant population movement into areas that were already poor and vulnerable before the influx[7] made it urgent to create an evidence base for policies to benefit Rohingya and host communities in Cox's Bazar. For the displaced, there was interest in understanding and tracking how key welfare indicators (such as food security, labor market engagement, and mobility) evolved as the crisis unfolded into the medium and long terms. For hosts, the large population influx was expected to have significant and heterogeneous impacts on multiple dimensions of welfare. However, evidence on this front was very limited due to lack of representative data after the influx.

The Cox's Bazar Panel Survey (CBPS) was designed in 2018 to produce a representative longitudinal survey to track key socio-economic and health indicators of Rohingya and the host Bangladeshi population in Cox's Bazar district.[8] The survey's main objective was to build an evidence base on the medium- and long-term impact of the Rohingya crisis on economic and social conditions among Rohingya and hosts. Due to the localized nature of the shock in the sub-districts of Ukhia and Teknaf, it was important to compare the

[2] United Nations High Commissioner for Refugees (UNHCR) refugee population factsheet (as of 15 July 2019) http://data2.unhcr.org/en/situations/myanmar_refugees.

[3] Staff calculation based on UNHCR refugee population factsheet (as of 15 July 2019).

[4] Staff calculation based on UNHCR refugee population factsheet (as of 15 July 2019) and Bangladesh Bureau of Statistics (BBS) population census 2011.

[5] Figures account for the total population of displaced Rohingya persons reported by the UNHCR (new and old arrivals).

[6] This is the density reported for Camp 3 in Ukhia in the UNHCR assessment "Settlement and protection profiling of all camps Ukhia/Teknaf, Cox's Bazar, Bangladesh, Round 5, July 2019."

[7] Host households in Cox's Bazar live in largely rural areas of the district, where consumption poverty is both high and significantly worse than the national average of 24.5 percent. Poverty rates outside the district capital of Cox's Bazar Sadar are higher than both the national and district averages, with the primary hosting subdistricts of Teknaf and Ukhia reporting small-area poverty estimates of 30 percent and 40 percent respectively. In 2016, only 55 percent of adults over 18 in Cox's Bazar reported being literate. In addition, only half of all households reported access to electricity, and less than 3 percent reported access to piped water. According to the 2011 census, infrastructure and social indicators are also significantly worse outside the district capital and notably in the primary hosting sub-districts.

[8] The CBPS is a partnership between the Yale Macmillan Center Program on Refugees, Forced Displacement, and Humanitarian Responses (Yale Macmillan PRFDHR), the Gender & Adolescence: Global Evidence (GAGE) program, and the Poverty and Equity Global Practice (GPVDR) of the World Bank.

situation between areas that were closer to the location of the large Rohingya camps and those areas farther away in the district. Accordingly, the sampling strategy for the CBPS had to meet two requirements: (i) be representative of both recently displaced Rohingya and host community households in Cox's Bazar; (ii) stratify these population groups based on other key characteristics, such as the prevalence patterns for the camp and non-camp Rohingya displaced populations across Cox's Bazar district.

3. Sampling Strategy

Approach to define the survey strata

Determining the prevalence of Rohingya displaced in camps and non-camp areas

In this particular crisis, a large share of the Rohingya displaced located in camps. Key population statistics for camp-based populations were both frequently updated and made publicly available, and population quick counts in camps were implemented approximately every three months by the International Organization for Migration (IOM). These population counts were complemented by United Nations High Commissioner for Refugees (UNHCR) data on new arrivals. Together, the IOM and UNHCR data suggested that a significant share of Rohingya who had arrived during the 2017 influx continued to live in camps in the Ukhia and Teknaf sub-districts more than a year after the onset of the crisis**.**

However, largely porous camp boundaries[9] and evidence from survey data[10] collected in the year after the crisis suggested that Rohingya displaced were not fully confined to camps. While some work opportunities could be generated within camps, data indicated that many Rohingya were either leaving and returning to camps daily for work purposes or else relocating entirely outside camps to pursue work.[11] For the purposes of obtaining a representative survey of Rohingya displaced it was important to determine the true prevalence of displaced Rohingya outside camps.

Due to the absence of a comprehensive sampling frame for Rohingya outside camps, IPA Bangladesh had conducted a quick population count exercise which provided an opportunity to assess the likely prevalence

[9] Based on satellite imagery, anecdotal evidence from multiple local actors, and field observations.

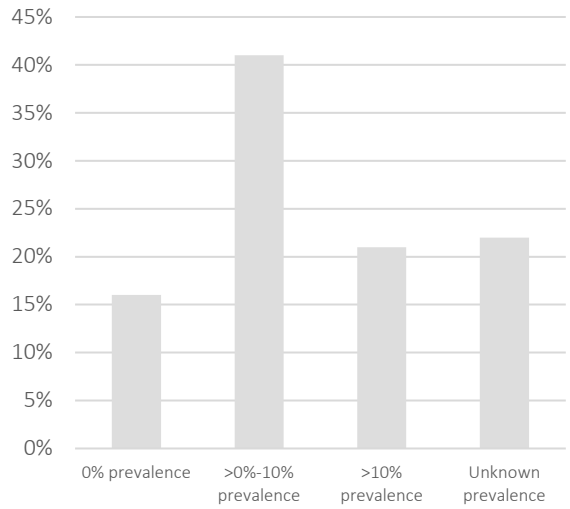
[10] The December 2017 Rohingya Emergency Vulnerability Assessment (REVA) and June 2018 Community-Based Monitoring System (CBMS) surveys, administered by the World Food Programme (WFP), found that over half of Rohingya displaced reached by telephone surveys reported working, despite a formal ban on work for Rohingya displaced (World Bank 2019).

[11] It was evident from field observations and reports that camp boundaries were largely porous. In addition, other survey data collected in the year after the crisis suggested that Rohingya displaced were not fully confined to camps. This was further corroborated by data collected by the World Food Programme as part of the Rohingya Emergency Vulnerability Assessment (REVA) in December 2017 and the Community-Based Monitoring System (CBMS) surveys in June 2018. Over half of Rohingya displaced reached by phone surveys for these assessments reported working, despite a formal ban on work for Rohingya displaced (World Bank 2019). These data suggested that some Rohingya displaced were either leaving and returning to camps every day for work, or else relocating entirely outside camps to pursue work opportunities.

of Rohingya outside camps. The IPA population count consisted of a census of unions in seven upazilas of Cox's Bazar district, which accounted for 76 percent of all unions in Cox's Bazar district.[12]

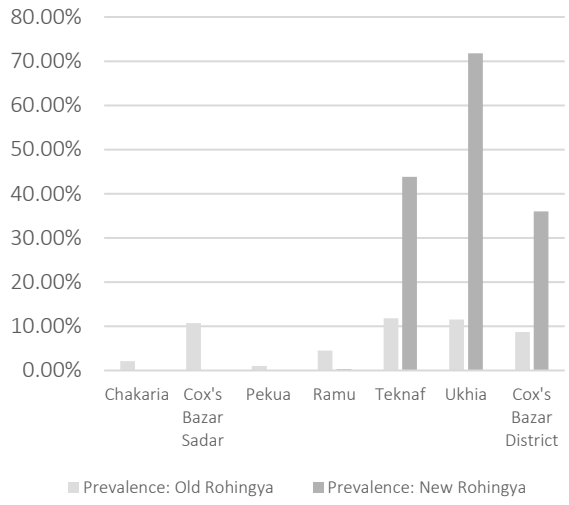
According to the IPA count, 57 percent of unions in Cox's Bazar reported a refugee prevalence of less than 10 percent, with 21 percent of unions reporting a prevalence of more than 10 percent. Unions reporting a positive refugee prevalence tended to be larger, covering about 11 villages on average. Figures 1 and 2 present figures on refugee prevalence outside camps at the union and upazila levels. Prevalence is reported starting from the union level (instead of the village), as this was the lowest unit for which complete refugee and host information was available.

Figure 1: Distribution of unions in Cox's Bazar by Rohingya prevalence



Source: 2011 census for host population count; IPA population count for Rohingya displaced.

Figure 2: Prevalence of Rohingya displaced by upazila

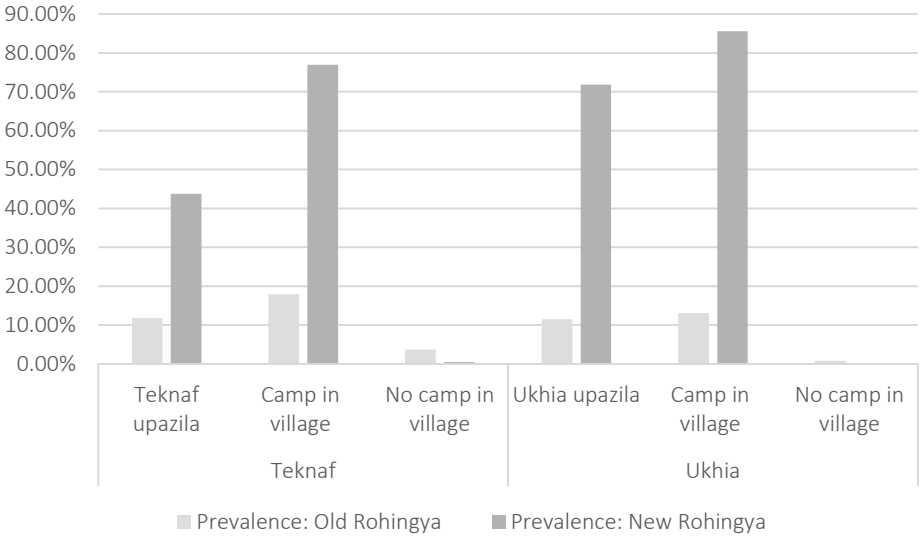


Source: IPA Population Count, 2018. Old Rohingya refers to displaced population arrived before September 2017. New Rohingya refers to population arrived after September 2017.

At the upazila level, Rohingya prevalence was low outside Teknaf and Ukhia, the two upazilas that contain all the camps. Cox's Bazar Sadar upazila, which is the district capital, was the only upazila where Rohingya prevalence exceeded 5 percent; almost all of these cases came from the older cohort of Rohingya displaced (Figure 2). In Ukhia and Teknaf, which reported a significant Rohingya presence, displaced households seemed to be located almost entirely in camps. In particular, villages with no camps reported prevalence rates of 0.8 and 4.2 percent respectively (Figure 3).

[12] The IPA count excluded Kutubdia and Maheshkhali upazilas where refugee prevalence was expected to be low as there were islands.

Figure 3: Average prevalence of Rohingya households, for Teknaf and Ukhia by camp and non-camp villages



Source: IPA Population Count, 2018. The IPA population count counts with 201 villages in Teknaf and 209 villages in Ukhia. Old Rohingya refers to displaced population arrived before September 2017. New Rohingya refers to population arrived after September 2017.

In summary, analysis of the IPA population count indicated a very small presence of recently arrived Rohingya outside camps and in upazilas that did not host camps. The few Rohingya who were reported as living outside camps were located in the adjoining upazilas of Cox's Bazar Sadar and Ramu that border Ukhia. These persons were almost all from the older cohort of Rohingya who arrived before the 2017 influx.

However, analyses of the count highlighted two challenges of relying on these data alone to conclude that the prevalence of Rohingya displaced outside camps was low. First, due to inconsistencies in the use of geographic codes, 35 percent of the villages covered by the IPA count could not be matched to the official Housing and Population Census host population data, and therefore no refugee prevalence information was available for these villages.[13]

Second, the IPA count relied on a non-standard methodology for collecting population numbers. Standard quick counts are typically used by all national households' surveys and even population censuses, to inform segmentation into census tracts or enumeration areas (EAs). Similar counts are also done for refugees by the International Organization for Migration (IOM). The reliability of those quick counts relies on getting

[13] Bangladesh has three administrative tiers below the district: upazilas, unions, and mauzas. Census villages, which are not an administrative unit, come below the mauza (a mauza in Cox's Bazar includes, on average, 3.9 villages). The IPA count recorded only two out of the three administrative levels and did not use census geocodes for census villages. This resulted in matching issues and led to a significant challenge in generating complete data on refugee prevalence at the village level, and in validating the count based on other auxiliary data sources before using it as a reliable sample frame.

estimates from a large set of key informants identified within a geographical unit that is small enough to be able to provide an estimate with low measurement error. In other words, quick population counts are usually done at the lowest feasible geographic unit, as informants are more likely to provide accurate information on units that cover smaller populations. The standard practice for implementing such a count in Bangladesh - followed both by the censuses and the IOM's count of Rohingya displaced in camps – focuses at the village/camp-block level. Camps and blocks cover, on average, 359 and 109 households each in Cox's Bazar district (Table 1). The IPA quick count, however, was implemented at the union level using a small set of union-level legislators as key informants. While the union is the lowest administrative level that has an office, the size of the population included in a union does not lend itself to an accurate quick count. This is particularly true for Rohingya displaced, for whom no administrative data are available.

Table 1: Average population by administrative and geographic units

	Number of units	Average number of persons	Average number of households
Hosts			
Unions	110	20,484	3,758
Mauzas	339	6,647	1,219
Villages	1,150	1,959	359
Rohingya displaced			
Camps	34	26,739	6,271
Blocks	1,953	465	109

Source: 2011 Bangladesh census and IOM NPM Round 12. Note: Averages for Cox's Bazar district

On average, a union in Cox's Bazar covers 10.5 villages and 3,758 native/host households. The IPA count asked a small set of union-level key informants [14] to provide village-level data on refugee counts, disaggregated by "old" and "new" Rohingya displaced. This approach was not likely to generate accurate numbers on prevalence. Moreover, the inaccuracy was most likely to pose problems for the key population of interest in this listing: Rohingya displaced living in villages outside camps.

Given these concerns, we implemented a full household listing in a random sub-sample of IPA count villages to validate the IPA counts. These villages were stratified for refugee prevalence based on refugee counts in the IPA listing and population counts from the 2011 census. Three types of prevalence were defined based on these data, to form the strata for the validation household listing (Table 2). The validation listing was done in 33 villages randomly chosen from the strata without refugee camps in November 2018. On average, one village was chosen from each mauza, across six out of the seven upazilas in Cox's Bazar district.

[14] The key informants were government officials in union level offices, which are the lowest local government administrative units in Bangladesh. IPA went to the respective union offices and asked for population counts.

Table 2: Prevalence strata for validation listing

Prevalence	Unions	Census HH	Rohingya HH out of camps	Camps	Rohingya HH in camps
Prevalence= 0	42	122305	0	0	0
Prevalence<.1	45	193626	7508	0	0
Prevalence>.1	18	59445	10251	0	0
With camps	5	38026	3571	27	281150
Total	110	413402	21330	27	281150

Source**:** Union-level prevalence using the number of local households (based on the 2011 census) and the number of out-of-camps Rohingya (based on the IPA population count). Note: HH = household.

This validation exercise implemented a full household listing. This involves going to each household in the chosen villages with a set of 12 questions that included: recontact details (household head name, phone numbers); demographic information on adolescents; and questions related to how long the household had been at that location, as well as details on where they had come from. If a village had a total population of fewer than 500 households, the full village was listed. This was the case for 27 out of the 33 villages. In the remaining six cases, the villages were divided into segments of 100-120 households each, and three segments were randomly chosen for the listing. If key informants at the village level indicated that the Rohingya households lived in particular segments, then one such segment was chosen as one of the three segments where a full listing was to be implemented. Additional segments indicated by the informant might also be included for full listing through random selection.

Table 3 summarizes Rohingya prevalence by upazila. In addition to a low average prevalence of 1.3 percent in the validation sample, a significant proportion of villages reported zero prevalence. Figure 2 and Table 3 let us compare the refugee prevalence based on the IPA population count and the household listing in the validation sample. The full household listing reported different numbers compared to the key-informant-based IPA population count, both for the number of households and refugee prevalence. However, the full listing confirmed the low prevalence of Rohingya displaced outside camps.

Table 3: Prevalence by upazila, validation household listing

Upazila Name	Number of Villages	HH on the validation	Prevalence	Average prevalence within villages	Number of villages with 0% prevalence
Chakaria	6	1,630	1.10%	0.54%	4
Cox's Bazar Sadar	10	4,221	2.99%	2.03%	4
Pekua	2	501	0.00%	0.00%	2
Ramu	3	508	2.95%	3.16%	1
Teknaf	7	2,830	1.41%	1.36%	3
Ukhia	5	1,467	0.14%	0.08%	4
Total	33	11,157	1.80%	1.30%	18

Source: Validation household listing, 2018

This two-step data collection on Rohingya refugee prevalence confirmed that prevalence in host communities was low, and that this was the case not only for newer Rohingya displaced, but for the older cohort of displaced, as well. These figures could still to some extent reflect the displaced population's mobility, as well as Rohingya's unwillingness to reveal their identities to surveyors. However, the design of the validation listing questionnaire sought to minimize this effect by asking a series of questions that related to how long the household that been resident in the host community, rather than direct questions on respondents' self-reported identity as Rohingya. Based on the validation findings, the final decision taken was to use administrative population data available from IOM NPM.

This pattern of refugee prevalence supported having one stratum for the Rohingya displaced living in camps. The sampling strategy for the CBPS therefore focused on generating representative estimates for the camp based Rohingya population in Cox's Bazar district.[15]

Stratification of the host population in Cox's Bazar district

For hosts, the sampling strategy was designed to account for the differential implications of a camp-based concentration of close to a million Rohingya displaced for different areas of Cox's Bazar. As in other displacement settings, the welfare among members of the host population is likely to vary with distance to camps. Hosts living closer to camps may be more likely than those living farther away from camps to compete with the Rohingya in labor and goods markets, as well as markets for natural resources like land and fuelwood. On the other hand, host households closer to camps may gain from camp-based economic opportunities,

[15] While the original sampling strategy was designed to be representative of all camp-based Rohingya displaced, campsites with older Rohingya displaced refused to participate in the listing due to other political sensitivities. This refusal was maintained despite many attempts. Since the older Rohingya displaced were not a separate stratum, a decision was made to drop these households from the survey.

compared to hosts living farther away. To distinguish between host communities that are differentially affected by the arrival of the Rohingya, the CBPS sampling strategy used a threshold of three hours' walking time[16] from a campsite to define two survey strata: (i) host communities with potentially high exposure (HE) to the displaced Rohingya, and (ii) host communities with potentially low exposure (LE).

Sample distribution among strata

The decision on how to distribute the sample size across the three strata faced the classical dilemma of whether to distribute across strata according to their population shares (which would deliver nearly optimal estimates for Cox's Bazar as a whole), or to allocate the same sample size to each stratum (which would deliver estimates of nearly the same quality for each stratum). Since both considerations were important, a 50/50 equal/proportional allocation between host and Rohingya displaced was chosen, following Aguilera et al. (2020). In other words, half the sample size was allocated to Rohingya displaced and half to hosts. Host communities were further distributed using a 50/50 equal/proportional allocation between high- and low-exposure communities. This procedure yielded 398 survey clusters: 200 clusters of refugee camps and 198 clusters of host communities equally distributed between high- and low-exposure areas. A full household listing was implemented in these clusters, and 13 households were then selected at random from each cluster for the survey. Table 4 shows the sample size at the design stage, as well as the realized sample size.

Table 4: Intended and actual sample size by strata

Stratum Code	Stratum Name	Distance to site (km)	Number of mauzas	HHs (census)	Pop (census)	Number of intended clusters	Number of actual clusters	Number of intended households	Number of actual households
0	Sites	Zero		213,198	909,120	200	192	2600	2496
1	High Exposure	0.1 to 15	33	83,482	463,528	99	99	1287	1287
2	Low Exposure	15.1 to +45	33	256,776	1,381,963	99	96	1248	1248
Total			**66**	**553,456**	2,754,611	**398**	**387**	**5174**	**5031**

Note: HH = household. Each cluster includes 13 households

Sampling Rohingya displaced households

For camps, high-resolution satellite imagery, together with population and household count data generated by the humanitarian agencies (UNHCR and IOM), provided the base for selection of EAs and the refugee sample. The sampling frame used for Rohingya displaced in camps was publicly available population counts from the IOM's Needs and Population Monitoring (NPM) Round 12 Site Assessment, which provided the most recently updated refugee population data at the time of survey design.[17] The data available identified 34

[16] Walkable roads and streets were mapped using OpenStreetMap (OSM) road networks. Three hours' walking time roughly equals 15 kilometers.

[17] NPM Round 12 was completed in October 2018, and the sampling strategy for this survey was finalized in December 2018.

camps comprising a total of 1,953 majhee blocks.[18] The camps were spread across 10 mauzas and five unions in the two sub-districts of Ukhia and Teknaf. [19] The average number of households within the 1,953 blocks was 103.

The survey was designed to visit 200[20] clusters of 13 households each in the refugee camps, and a two-stage stratification was implemented to identify the refugee sample. In the first stage, blocks (enumeration areas for camps) were selected for a full household listing, using a probability of selection proportional to the size of the block. After selecting the block, a full listing was carried out in each block as will be explained in the next subsection. In the second stage, 13 households were randomly selected from the full household listing.[21] For the adult module, a third stage was implemented within each household, in which two adults[22] were randomly selected from all the adults living in the household.

Selection probability for a Rohingya displaced household

Given the sampling design discussed in the preceding sections, P_k^R is the probability of selecting a refugee household (R) in an enumeration area k:

$$P_k^R = \frac{z \times n_k^R}{\sum_k n_k^R} \times \frac{1}{l_k^R}$$

Where,

z is the target number of enumeration areas to be selected (target of 200 blocks)

n_k^R is the number of refugee households in enumeration area k according to IOM NPM R12

l_k^R is the number of listed refugee households in enumeration area k (selection target of 13 households per block)

On the right-hand side of the equation, the first term represents the probability of selecting the block k in the first stage, and the second term the probability of selecting the household in the second stage.

[18] In the Rohingya community, local leaders are referred to as "majhee(s)." A majhee block thus refers to a collection or subset of households within the same area in a camp that are overseen by the same local or community leader. In the absence of a standard division of camps into sub-blocks across all 34 camps, IOM NPM executed population counts using majhees as their focal points, and the areas governed by individual majhees as the lowest geographical units.

[19] Bangladesh is divided into eight divisions (bibhag) and 64 districts (jela, zila, zela). For the purposes of local government, the country is divided into upazilas (sub-districts), union councils (or rural councils), and then mauzas. Urban areas are divided into municipalities or city corporations and then into wards. The administrative structure of Bangladesh thus includes: Division (admin 1) >> district/zila (admin 2) >> sub-district/upazila (admin 3) >> union/municipalities (admin 4) >> mauza/wards (admin 5).

[20] During the listing, eight blocks were not listed due to collective refusal by households.

[21] The survey covers at least one block in all the camps except Camp 20 Ext (CXB-234).

[22] An adult is defined as any household member aged 15 years or older.

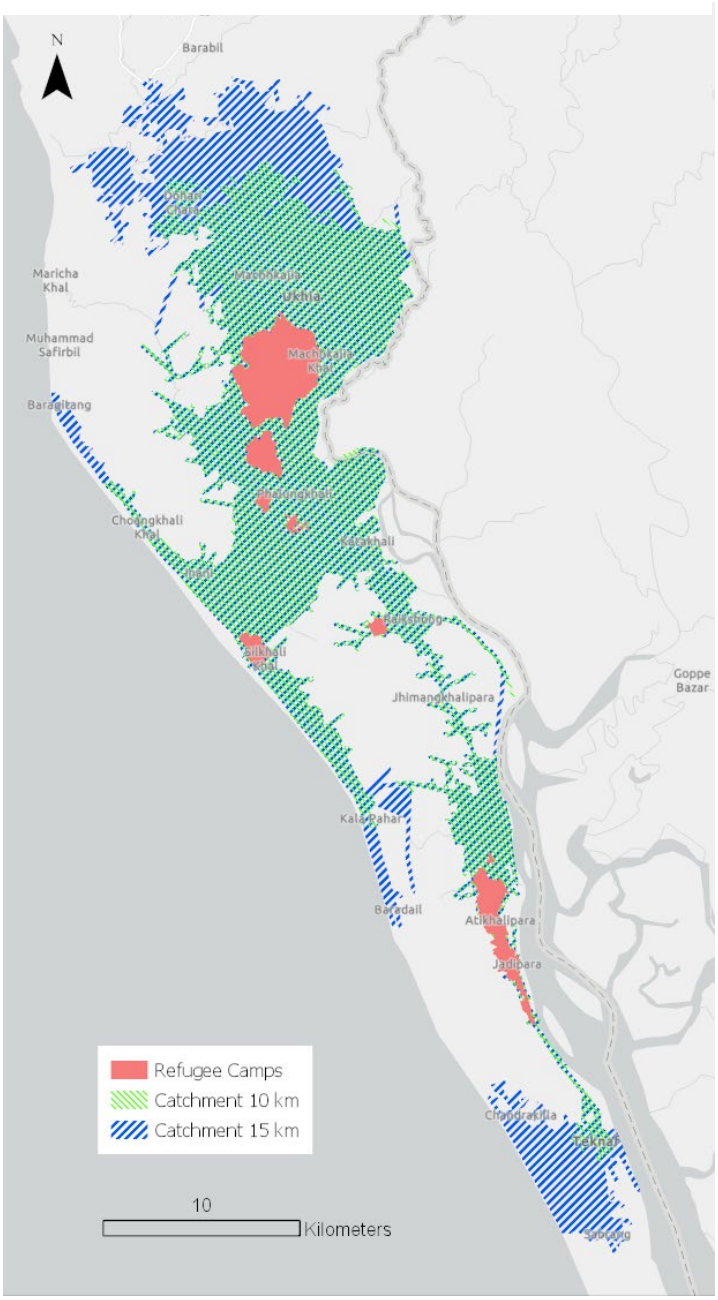
Sampling host households

Defining high-exposure (HE) and low-exposure (LE) areas

For the host population, a key sampling challenge was to define population strata based on differential levels of exposure to the refugee influx. As discussed above, the chosen sampling strategy established two strata for hosts in Cox's Bazar and portions of the adjacent Bandarban district: "high exposure" (HE) and "low exposure" (LE) strata. Households in HE communities were expected to interact more intensively with displaced Rohingya than households in LE host communities.

HE and LE host areas were defined using the road network information available in OSM. [23] Initially, various scenarios were developed for areas encompassed within a walking distance of from two to five hours from a refugee campsite. [24] Map 1 shows areas that were located within two (green) and three (blue) hours' walking distance from camps. Public transportation for long distances was both irregular and slow for hosts. Moreover, it was routinely checked at multiple check-posts on the main highway connecting Teknaf and Ukhia with Cox's Bazar town, thereby not offering a reliable way for Rohingya to leave camps either for daily work or to relocate. Camps were therefore not easily accessible to hosts living further away from camp sites, while most

Map 1. Catchment areas for 2 and 3 hours' walk from refugee camps

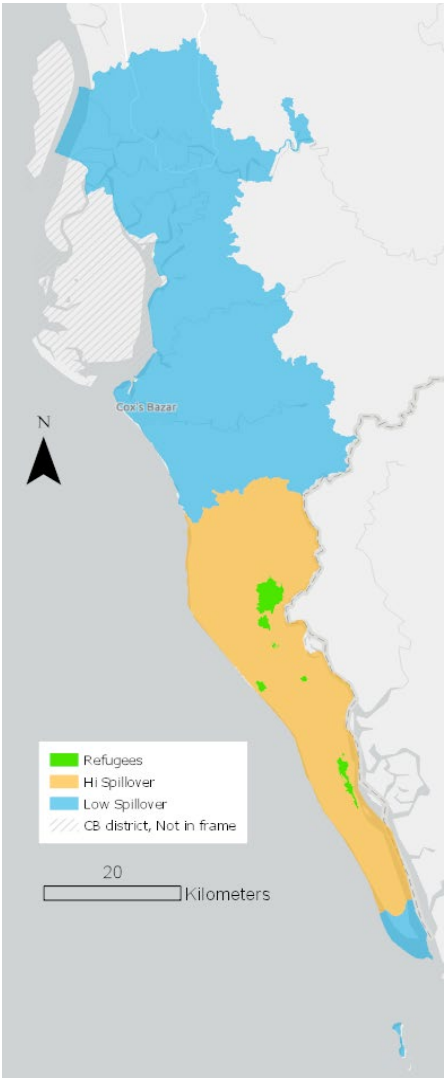


[23] Map data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org. OpenStreetMap contributors (2019).
[24] Assuming a constant walking speed of five kilometers per hour.

movement of Rohingya, if it occurred, was by foot. Given these conditions, it was decided to specify the high-exposure host stratum to include all mauzas[25] within a three-hour walk from a campsite (Figure 4).[26]

Demarcating enumeration areas within mauzas

Figure 4: CBPS strata



A combination of an outdated sampling frame, the low resolution of publicly available satellite imagery[27], and limitations on access to Bangladesh's 2011 census data posed additional design challenges for the host sampling strategy. The sampling strategy was originally intended to use 2011 census tracts or EAs.[28] However, using the census EAs was not feasible, due to limitations in access to the cartography required to locate the census EAs within census villages. This inability to geographically locate the EAs from the 2011 census meant that the 2011 census EAs could not be used as a sampling frame for the CBPS survey.

The smallest geographic unit for which digital maps were available was the mauza, which is also the lowest administrative level of government in Bangladesh. On average, a mauza includes more than 1,200 households. In 2011, census EAs had, on average, one-tenth of the population of a mauza; Cox's Bazar district had a total of 3,352 EAs, with an average of 123 households per EA. Mauzas were therefore significantly larger than EAs, and mauza outlines indicated in publicly accessible maps did not provide sufficiently granular demarcations for the team to delineate and sample EAs from this source directly.

To address these challenges, we constructed new enumeration areas for this survey. These EAs were constructed using publicly available

[25] A better approach to this situation would have been to choose only the census EAs within the areas designated as high exposure using the road network. Since this was not feasible, analysis proceeded at the next available level, mauzas. Mauzas were the lowest administrative level with shapefiles available.

[26] The high-exposure area includes three mauzas in Bandarban, to be consistent with the three-hour walking distance criterion. The low-exposure stratum included the rest of the mauzas in Cox's Bazar district and four mauzas in Bandarban district that were within 45 kilometers' walking distance from the camps. This survey is not representative of Bandarban as a district.

[27] OpenAerialMap.org.

[28] These census EAs are used as the primary sampling unit in all national household surveys in Bangladesh, including the most recent national Household Income and Expenditure Survey (HIES), implemented in 2016/17. Using these EAs as the primary CBPS sampling unit would allow CBPS data to be used along with data from national surveys like the HIES and the national census, and would enable comparisons with previous representative household surveys for hosts.

data sets and based on the systematic field validation protocol. These data and protocols are described next.

The geospatial and census data available to segment the mauzas into EAs was very limited. For geospatial data, we had access to mauza shapefiles from the Bangladesh Bureau of Statistics and publicly available Google Earth imagery. While the best practice for segmentation would have been to implement a roof-counting exercise and use geographical boundaries visible on the imagery to create these segments, the resolution of the imagery available was not sufficiently high for such an exercise to be reliable. To address this problem, we developed an innovative algorithm to segment the selected mauzas. The approach used building outlines visible on OSM as of December 2018 (Figure 5.A), combined with mauza shapefiles and data from the 2011 census. Satellite or drone imagery (when available from IOM or Google Earth Pro) was used to visually validate the OSM building information.

Figure 5.A: OSM Building demarcated

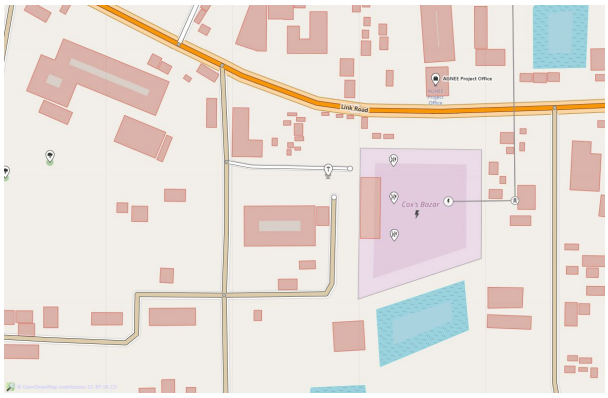


Figure 5.B: OSM Buildings demarcated with Satellite imagery.

Figure 5.C: Mauza Chowdhury para automated segmentation result

Figure 5.D: Mauza Chowdhury Para randomly selected segments for listing.

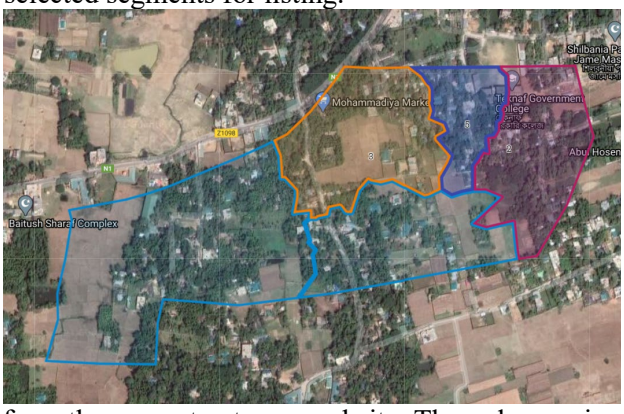


Figure 5.A and 5.B are snapshots a Cox's Bazar area from the open street map website. The polygons in red are the ones used in the segmentation process. Figure 5.C and 5.D is the resulting segmentations of Mauza Chowdhury Para and the maps used by the field teams in Google maps.

The process adopted to generate the enumeration areas was conducted in the following sequence. First, Mauzas selected in the first stage were identified. Next, within the mauzas, any structure smaller than seven

square meters or larger than 500 square meters was eliminated. The lower bound was obtained using data on average housing size in rural areas from the Bangladesh Integrated Household Survey (2015). The available imagery was contrasted with the buildings obtained from OSM to ensure reasonable coverage (Figure 5.B).[29]

We took the number of households in a mauza (according to the 2011 national census) and calculated how many segments would need to be created to obtain 100 households per segment. At the same time, the number of buildings in the segment was calculated. Given these two numbers, it was assumed that households were uniformly distributed in the buildings available in the data set. The team then proceeded to calculate the number of buildings needed per segment to obtain the same number of segments as if segments had been defined based on population.[30] This procedure was undertaken to connect the number of households in the 2011 census data to the buildings observed in the OSM data set. If the underlying assumption was reasonable, dividing the segments using buildings would produce segments with roughly the same number of households[31] within a particular mauza.

Using the centroids of each building, the team generated non-overlapping polygons within the mauza (Figure 5.C). To generate the polygons, the team minimized the total squared distance from the center of the segment to each building in the segment, given the minimum and the maximum number of buildings that are allowed in each subset.[32]

The segments were then checked, to solve any inconsistencies due to irregular shapes of particular mauzas. Once this process was complete, the team moved forward with a ground-truthing exercise to verify assumptions, followed by listing of the selected segments (Figure 5.D).

This process described had to accommodate two detours: First, for some rural northern mauzas the availability of demarcated buildings in OSM data was low, so in those cases a different approach was used. Since enumeration teams would in any case be sent to the field to list, segments were created based on the assumption that the population was uniformly distributed across the area. Enumeration areas were then created that field teams would be asked to list using field protocols.

Second, the mauza shapefiles less accurate in urban areas: In general, the shapefiles available were not a perfect match to the actual areas; with this in mind, the team proceeded with listing in the selected mauzas using the available boundaries. In two cases (Light House Para and Nazirar Tek) in Cox's Bazar Sadar, the location of the mauzas on the shapefile did not coincide with the actual location of the areas.[33] The field team

[29] In 13 cases the coverage was low, so a different approach was adopted to segment within those mauzas.
[30] For example, in Shahparir Dwip, the 2011 census reports 4,351 households, and 2,081 buildings were found. To obtain 100 household segments, it was necessary to create 44 segments. Assuming that the distribution of population is uniform, each segment will have around 47 buildings (2081/44).
[31] In a sense, the listing process was able to validate the assumption, which worked surprisingly well in rural areas.
[32] Although a set number of segments had to be created, this process needed to run with some flexibility. The number of buildings had an upper and a lower bound. For example, if 200 buildings are observed and two segments are needed, the method will use 99 as a lower bound and 101 as an upper bound as a first try, then proceed until a solution to the problem is obtained and the two segments are drawn. On average, the segments in that area will have the same number of buildings. For more details on the process, see https://pro.arcgis.com/en/pro-app/tool-reference/geostatistical-analyst/generate-subset-polygons.htm.
[33] For example, Light House Para in the shapefile was located on the beach with no households, whereas in actuality, the area is in a moderately populated area much farther inside the city.

redrew the boundaries for those two places, and the segmentation was done again following the process previously described.

The last challenge with using this segmentation method was that buildings demarcated on OpenStreetMaps do not typically correlate directly with population figures due to factors such as the lack of information on building heights and data being collected through public reporting. Subsequently, a ground-truthing component was designed in order to validate the mauza segmentation. The ground validation exercises yielded several key findings, notably:

- The segmentation of EAs was relatively less accurate in densely populated urban regions as opposed to rural mauzas with scattered settlements.
- The number of households found within these segments varied, with some segments having more households than expected and some having fewer than expected.

Field protocols were designed in order to address these concerns. The protocols involved combining GPS coordinates and also training enumerators in the use of Google Maps to allow for systematic field-based adjustment.

Drawing the host sample

Once the strata were defined, the first stage of selection was done at the mauza level by strata. The selection process used probability proportional to the population size of the mauza. As Table 4 shows, the sample for host communities was divided equally between the two strata, so 33 mauzas were selected in each stratum. Within each selected mauza, three segments were designated with equal probability of selection. Mauzas were allowed to be selected more than once.[34] If this occurred, then the number of selected segments within the mauza increased accordingly. For example, if a mauza was selected twice in the sample, six instead of three segments would be selected from the mauza, each segment containing 100-150 households.

The second-stage selection was made using the enumeration areas created for the selected mauzas. All the segments within a mauza had the same probability of being selected. The field team listed all households within the boundaries of the selected segment, subject to field protocols where necessary.

With the listing data available, the third stage was to select 13 households in each of the listed segments, with every household having the same probability of being selected. The selected households were to be interviewed.

Selection probability for hosts

[34] Since the selection was done using probability proportional to the population size of the mauza with replacement, bigger mauzas had higher probability of being selected more than once.

Given the sampling design discussed in the preceding paragraphs, P_{ijk}^{H} is the probability for selecting a host household (H) of stratum i, in mauza j, and enumeration area k.

$$P_{ijk}^{H} = \frac{m_i \times n_{ij}^{H}}{\sum_k n_{ij}^{H}} \times \frac{1}{n_{ij}} \times \frac{1}{l_{ijk}^{H}}$$

Where:

m_i is the number of mauzas selected in each stratum i.

n_{ij}^{H} is the number of host households according to the 2011 census in mauza j of stratum i.

n_{ij} is the number of segments/enumeration areas in mauza j of stratum i.

l_{ijk}^{H} is the number of listed households in stratum i, mauza j, in enumeration area k (the target was to select 13 households per segment).

The three terms on the right-hand side of the equation respectively represent the probability of selecting a mauza j within a stratum i in the first stage; the probability of selecting an enumeration area k in a mauza j in the second stage; and the probability of selecting a specific host household in the third stage.

Listing methodology

The main listing exercise was an important part of the preparation of the survey as existing population counts for the host population dated 2011. Analysis of the IPA count for the host population indicated that this exercise did not provide reliable updated population counts of the host population due to the partial coverage of areas, as well as the large estimation error coming from key informants reporting figures centrally. Rohingya counts in camps were updated more frequently, however, movements across areas or campsites could have happened between the IOM count and the data collection. Therefore, a listing exercise was conducted for both host and Rohingya displace in the selected enumeration areas. In this section we explain how the listing exercise was conducted and highlight some challenges that arose when implementing the listing and their solutions.

The resources available for refugee camps and host communities were significantly different, but the proposed listing methodologies for both samples were based on the respective maps available for the enumeration areas. Field teams were instructed to identify the enumeration areas with assistance from local authorities and using landmarks from the maps. Once at the enumeration area, the listing instructions were divided into three categories:

- For EAs with less than 400 households in both camps and hosts, the entire EA was to be listed.
- For larger EAs with more than 400 reported households in camps (based on NPM population counts), the block was divided into sub-blocks. The segmentation methodology yielded sub-blocks of 100-150 households each based on the number of buildings and geographical

boundaries visible in the drone imagery. Four sub-blocks were then randomly selected and fully listed.

- For larger EAs with greater than 400 households in hosts, a fanning protocol was developed for listing based on a ground-truthing exercise described in detail below.

Ground truthing and developing the listing methodology for camps

For the Rohingya camps, we had access to publicly available high-quality drone imagery from IOM's Needs and Population Monitoring program for the Round-12 data that had been used as the sampling frame. The maps in theory could allow us to geographically locate the exact 200 blocks that had been sampled for the survey. In addition, the field team intended to use the IOM-assigned Block IDs (6-digit codes) and the Local Block Names specified in the NPM data to identify blocks inside the camps.

The field team was instructed to go to the campsite management office and ask staff or volunteers for assistance in locating the sample blocks in the camp. Printed IOM drone maps were provided to navigate inside the camps using prominent buildings and other geographical landmarks.

The first challenge faced in the field was inconsistencies in block names used across the camps. The field team had mixed experiences in being able to identify the correct blocks using the NPM Block IDs and Local Block Names provided. Some camp authorities were able to identify blocks using the IOM-assigned IDs, and some were identified using the local block names. However, in some camps neither of these options was feasible. This was explained by the following circumstances:

- The IOM NPM data used majhees[35] as their main point of contact and did the block mapping based on "majhee boundaries." The majhee governance system is an informal system, and thus it was realized that these assigned boundaries and local names are relatively fluid and subject to change.
- Half of the camps in Ukhia and Teknaf were managed by UNHCR and the other half by IOM. Thus, while the NPM block IDs and names were familiar to IOM-governed site management, camp managements governed primarily by the UNHCR or the Relief and Repatriation Commissioner (RRRC)/army were not familiar with the information.

In addition, the field team also found it difficult to navigate the camps on the ground using the printed versions of the drone imagery, since they had no GPS support. After a validation exercise on the issues faced on the ground, the field protocols to identify sample blocks were revised.

The blocks successfully recognized by the respective camp management were fully listed. For blocks that could not be located with help from the camp management, the field team was given new instructions. The NPM data for the majhee blocks provided GPS coordinates for the centroid of each block. These coordinates were shared with the field team, who were instructed to locate the points using Google maps. Once they had

[35] As discussed in the main text, new Rohingya refugee camps (established after the influx of August 2017) are governed by local leaders called "majhees." A majhee block is defined by IOM as an area governed by one majhee.

located the center of the assigned block, they had to identify the majhee who was in charge of the majority of households surrounding the point and list all households under said majhee.

UNHCR provides a publicly available camp navigation application that is primarily used by site management and NGOs operating inside camps. A customized version of the tool, containing the NPM R12 majhee mapping system and outlining the CBPS sample blocks, was shared with the team in the second half of the camp listing period. After the tool was received, the teams used the application to locate and list within demarcated sample blocks.

Ground truthing and developing the listing methodology for hosts

The host mauza segmentation was conducted digitally using an algorithm that divided each host mauza into segments of approximately 100-150 households based on population counts from the Bangladesh Population Census 2011 and the number of buildings within mauza boundaries visible on OSM. The proposed listing methodology instructed field teams to locate the selected segments using maps generated on Google Maps and do a full listing within those segments. However, owing to the issues faced during the camp listing, the suboptimal quality of the imagery, and heterogeneous settlement patterns across host communities, it was judged necessary to carry out a ground validation exercise before the actual listing could start. Non-sample mauzas were selected for a two-day validation exercise which highlighted that the actual number of households in some of the selected segments was significantly higher or significantly lower than the expected 100-150 households. For these segments, a number of field protocols were developed in addition to the core listing guidelines:

- *For segments with fewer than the expected number of households (that is, fewer than 100 households)*, if the border of the segment cut through a settlement cluster belonging to another non-sample segment, the field team was instructed to continue listing along the border of the segment until they had listed approximately 100 households.
- *For segments with fewer than 100 households and with no avenue for expanding* listing into clusters along the segment borders, field teams were instructed to complete listing among the households inside the border. In the case where the selected segment had fewer than 10 households, the field team was provided with a replacement segment to list.
- *For segments with more than 150 household*s, a deviation of 50-60 households was accepted, and field teams were instructed to list the full segment. However, for segments with significantly more than 150 households, a fanning protocol was followed, where GPS coordinates were provided for the settlement clusters inside the segment, and each field team was instructed to list at every one of those clusters, starting from the coordinates provided and fanning out in all directions. In total, the field team was expected to list 150 households in a segment, combining all the marked clusters.

4. Discussion

The sampling strategies described in this paper were designed with the aim of producing a comprehensive longitudinal household survey of social, economic, and health indicators from a representative sample of Rohingya displaced and the host Bangladeshi population in Cox's Bazar district. This effort was important to set an infrastructure that allowed for the representative tracking of both hosts and displaced in later years. In particular, with the COVID-19 pandemic, the existence of this representative survey with detailed contract information for all respondents allowed to implement a series of phone-based follow-ups that were critical to inform about the impacts of the pandemic in Cox's Bazar.[36]

This paper describes the strategy implemented to design a representative survey of this nature, using a combination of census data and publicly available humanitarian and geospatial data. Two different data collection exercises were carried out to assess the prevalence of Rohingya displaced outside the camps to inform the sampling strategy for Rohingya displaced. Administrative data from humanitarian agencies were used to design the sampling frame within the camps. Drone imagery and digital maps were used to implement the listing within camps and host communities. Two different open-source data sets were used to inform the design of the host strata and help generate the host enumeration areas. Government data such as the 2011 population census and administrative shapefiles were also used.

This paper aims to share these different strategies and lessons learned for researchers interested in designing representative surveys in displacement contexts. This is particularly relevant when attempting to do such exercises close to the time the displacement occurs, as it is unlikely that national sampling frames are updated. The analysis also highlights the importance of listing as a way to update population counts, and some of the potential risks of key informant approaches to estimate population counts. We also highlight the value of humanitarian data registries to inform ex-ante selection probabilities in contexts where those displaced are largely located in camps or areas covered by those data.

Although it was possible to solve many data gaps in this case, it is important to keep in mind that institutional solutions will make sampling frames more credible and compatible across surveys. Efforts to maintain detailed census cartography and consistent geocodes across surveys are central. As well as steps such as including refugees and non-nationals in national sample frames, which can provide a more sustainable infrastructure to design more accurate surveys and achieve more effective policies.

[36] https://www.worldbank.org/en/country/bangladesh/brief/cox-s-bazar-panel-survey-briefs.

References

Aguilera, A., Krishnan, N., Muñoz, J., Riva, F. R., Sharma, D., & Vishwanath, T. (2020). "Sampling for Representative Surveys of Displaced Populations." In J. Hoogeveen & U. Pape (Eds.), *Data Collection in Fragile States* (pp. 129-151). Palgrave Macmillan, Cham, Switzerland.

UNHCR (United Nations High Commissioner for Refugees) (2019). Bangladesh: UNHCR Camp Settlement and Protection Profiling - Round 5 - July 2019. Retrieved from *https://reliefweb.int/report/bangladesh/bangladesh-unhcr-camp-settlement-and-protection-profiling-round-5-july-2019*

World Bank (2019). "Short-Term Poverty Impacts of the Rohingya Crisis in Cox's Bazar, Bangladesh." In *Bangladesh Poverty Assessment: Facing Old and New Frontiers in Poverty Reduction*, volume II (pp.377-394). Washington, DC: World Bank.