ML-based geographic sampling frames miss transitory populations in fragile regions

Andrea C. Caflisch, Daniel Masterson, Stephen D. O'Connell,
Ettan Patel, and Julia Smith-Omomo

May 15, 2026

Abstract

Accurate sampling frames are essential for representative surveys in post-conflict settings, yet standard methods — in-person listing and manual satellite digitising — are costly and slow. We document and evaluate a hybrid workflow that combines Microsoft's open-source machine-learning building footprint repository with contemporaneous high-resolution satellite imagery to construct sampling frames across 18 displacement-affected communities in Iraq. In our subsequent household survey, the approach achieved an 87% residential accuracy rate with fewer than 1% of ML-identified points requiring deletion. Using weighted linear probability models, we find that ML-retained and manually placed points yield significantly different respondent compositions in rural and urban communities. In rural areas, manually placed points capture a higher share of internally displaced persons (IDPs); in urban areas, they capture a higher share of returnees and a lower share of hosts. We provide evidence of two mechanisms driving these differences. First, algorithmic footprints miss tents and informal shelters, commonly used by IDPs in rural areas. Second, aerial and artillery bombardment during wartime reshaped the building stock in some urban areas. Our findings suggest that a mixed methodology can substantially reduce cost while preserving frame quality for the displaced and mobile populations that are hardest to reach and most important to count.

Keywords: machine learning; building footprints; sampling frames; survey methodology; displacement; internally displaced persons; post-conflict; satellite imagery; Iraq; humanitarian assistance

JEL Classification: C81, C83, D74, F22, O12

Contact Information. Caflisch: Department of Political Science, University College London and International Organization for Migration; Masterson: Department of Political Science, University of California, Santa Barbara; O'Connell: Department of Economics, Emory University; Patel: Department of Economics, Emory University; Smith-Omomo: International Organization for Migration. Corresponding author: O'Connell (soconnell@emory.edu).

1 Introduction

Post-conflict environments present severe challenges for the measurement of social and economic conditions, complicating efforts to target assistance where it is most needed. The breakdown of administrative systems, mass population displacement, and the rapid construction of informal settlements render pre-existing data — even when it exists — unreliable or obsolete. In such contexts, aid organizations, government statistical agencies, and researchers often construct sampling frames — enumerated lists of geographic points where populations reside — before deploying survey teams. When these frames are inaccurate or incomplete, the consequences extend beyond survey quality: vulnerable populations, particularly internally displaced persons (IDPs) and returnees, may be systematically undercounted. This can not only introduce bias in aggregate statistics, but disproportionate miscounting of these populations can change the understanding of need and the distribution of aid or recovery programming within communities. Survey methodologists term this phenomenon *coverage error*: the systematic discrepancy between the target population and the population actually enumerated by the sampling frame (Groves et al., 2009). When coverage error is non-random — as is predictably the case when displaced populations are concentrated in structures that standard enumeration algorithms fail to detect — the resulting statistics misrepresent the distribution of need, potentially misdirecting resources toward groups already adequately enumerated. The challenge of accurately enumerating displaced populations is further compounded by their mobility; IDPs and returnees frequently reside in camps, informal settlements, partially destroyed buildings, and construction sites — locations that standard enumeration algorithms are more likely to miss or classify as uninhabited locations.

The economic and social consequences of displacement are well documented. IDPs and returnees consistently exhibit persistent gaps relative to host communities across measures including household consumption, asset wealth, food security, labor market participation, and subjective well-being (Fransen et al., 2017; Kondylis, 2010; Verwimp and Muñoz Mora, 2018; O'Reilly, 2015; Admasu et al., 2021). Household- and community-level barriers to reintegration render some subgroups particularly vulnerable within returnee populations

(Fransen and Bilgili, 2018). The negative welfare impacts of displacement are often accompanied by differences in social cohesion outcomes, although the direction of the latter varies contextually (Tellez and Balcells, 2025; Ruiz and Vargas-Silva, 2025). Accurately capturing displacement-affected populations in survey data is therefore critical not only for the administration of humanitarian and recovery programming, but additionally for statistical and evaluative undertakings in post-conflict settings. Statistical invisibility may render these populations "illegible" to both state and humanitarian actors, trumping efforts to deliver targeted development and social protection interventions and potentially contributing to documented welfare gaps.[1]

Established approaches to sampling frame construction in low-resource settings fall into two broad categories: in-person listing (Aguilera et al., 2019), and GIS-based enumeration using satellite imagery, as demonstrated in epidemiological and health surveys across a range of fragile settings (Galway et al., 2012; Escamilla et al., 2014; Wagenaar et al., 2018; Lin and Kuwayama, 2016). Both are considered methodological standards, yet both carry significant limitations in post-conflict environments. In-person listing is logistically complex, costly, and time-intensive, and in aid-dependent contexts risks generating confusion around registration for social assistance. Satellite-based manual tagging reduces these type of field risks, but remains dependent on existing methodological tools and the vintage of available imagery.[2] These constraints motivate the search for approaches that can produce accurate, up-to-date sampling frames rapidly, at low cost, and without requiring sustained field presence prior to survey deployment. Table A1 in the appendix summarises the key operational trade-offs across these approaches.

Recent advances in machine learning (ML) offer a promising alternative. Freely available repositories of ML-generated building footprints — most notably Microsoft's GlobalML Building Footprints dataset — enable rapid, low-cost population sampling frames without initial field deployment. A growing body of work now takes such ML-derived layers as direct

[1] While an expanding literature has established a link between legibility, state capacity and access to state services (Scott, 1998; Lee and Zhang, 2017; Bowles, 2024), the relationship between illegibility and other dimensions of marginalization — such as forced displacement — remains an open area for investigation.

[2] Some previous approaches, such as spin-the-pen, have been largely abandoned due to well-documented geographic sampling bias (Lin and Kuwayama, 2016; Bauer, 2014).

inputs to population enumeration and survey design: Wardrop et al. (2018) and Weber et al. (2018) use satellite-derived building footprints for census-independent population mapping; Thomson et al. (2017) develops an open-source tool for generating primary sampling units from gridded population data that rests on ML building layers; Thomson et al. (2020) reviews 43 gridded-population surveys across 29 low- and middle-income countries, most of which treat the underlying ML or modelled layer as given; and Boo et al. (2022) and Neal et al. (2022) use building footprints as core inputs to high-resolution population estimation models without systematic validation of the footprint layer itself. The broader value of pairing satellite imagery with ML has been established convincingly in the development economics literature: Steele et al. (2017) and Yeh et al. (2020) show that such models can estimate household wealth with accuracy approaching that of traditional surveys in sub-Saharan Africa, and Aiken et al. (2022) demonstrates that ML-derived estimates can improve the targeting of humanitarian cash transfers. These results, however, are obtained in settings with relatively stable building stock and higher-quality training data. In post-conflict environments, the satellite imagery used to train building detection algorithms may predate the conflict by years, leaving the model poorly calibrated to the informal structures, recently reoccupied ruins, and improvised dwellings in which displaced populations disproportionately reside. At least two sources of bias are predictable in such settings. First, temporary and improvised shelters — tents, prefabricated units, makeshift structures assembled from scavenged materials — lack the rooftop signatures on which building detection algorithms are trained, rendering their occupants invisible to the ML layer. Second, wartime destruction reshapes the building stock: aerial bombardment and artillery fire damage or destroy roof structures, and the buildings that replace them during post-conflict reconstruction may differ in appearance, location, or footprint from the structures recorded in pre-conflict training imagery. Both channels predict that displaced populations will be systematically undercounted, but through distinct mechanisms and in distinct settings — a distinction with implications for where manual validation effort should be concentrated. The validity of any ML-derived sampling frame in these contexts therefore depends critically on whether its output can be audited and corrected against current imagery — the core question this paper addresses.

We document and evaluate a hybrid methodology for sampling frame construction deployed as part of a large survey data-collection project in Iraq over the period of 2024-2025. Our approach combines Microsoft's open-source ML building footprint repository (Microsoft, 2022) with contemporaneous high-resolution satellite imagery procured from Planet Labs to produce audited sampling frames across 18 communities ranging from rural farming settlements to dense urban neighborhoods. We use the ML layer as a starting point, manually validate and correct its output against up-to-date imagery, and supplement building locations with manually placed centroids for structures the algorithm failed to detect. This workflow extends prior IOM GIS-based sampling efforts (International Organization for Migration, 2021; Wiens et al., 2021) by systematizing the ML-assisted component and generating evidence on its performance.

Our central empirical contribution is a descriptive one: we document the extent to which building identification method — ML-generated versus manually placed — is associated with differential capture rates of hosts, IDPs, and returnees, the respondent types of greatest policy interest in our context. We make no causal claim that the sampling approach itself is the mechanism driving this association; the observed differences may partly reflect the geographic co-location of building types and population groups, since found points are concentrated in areas of recent construction where returnees and IDPs disproportionately reside. What the comparison establishes, regardless of mechanism, is the magnitude of the coverage error a practitioner would incur by deploying the ML layer without manual validation. Using exposure-weighted linear probability models, we find that ML-retained and manually placed points are associated with meaningfully different respondent compositions. Manually placed points are associated with a larger share of IDPs and a substantially lower share of hosts relative to the ML baseline. These aggregate differences mask important heterogeneity: while in rural areas the gap between the manual and ML frame is larger for IDPs, in urban areas returnees are under-represented in the ML frame (Table 4). The direction of the gap for returnees additionally reverses across communities with different conflict histories (Table 8). We trace these patterns to two distinct sources of ML detection failure: shelter informality, which renders tents and improvised dwellings invisible to the ML layer

and fully explains the rural IDP–Found gap, and post-bombardment reconstruction, which causes the ML layer to miss returnees in communities where wartime destruction reshaped the building stock. These biases persist a decade after the bulk of the fighting. The overall residential accuracy of our sampling frame was 87%, and fewer than 1% of ML-identified points required deletion during validation[3].

These findings carry practical implications beyond the Iraqi context. While most directly relevant for population monitoring in displacement settings, the approach has applications for demographic measurement, humanitarian policy delivery, and survey sampling across a wider set of highly dynamic demographic contexts. Conflict-affected areas undergoing rapid destruction and reconstruction, high-mobility settings where populations reside in tents or temporarily unoccupied structures, informal and squatter settlements, and post-disaster environments all share the core challenge of mobile, undercounted populations in settings where traditional enumeration is impractical. The methodology described here relies entirely on open-source or commercially accessible tools — QGIS, the Microsoft Footprints repository, and commercially procured satellite imagery — and is therefore replicable by researchers and practitioners with modest resources.

Within political science, our work contributes to efforts — by both quantitative and qualitative scholars — to document and correct biases affecting data generation processes related to migration, forced displacement and humanitarian emergencies (Landry and Shen, 2005; Shaver et al., 2025; Alrababah et al., 2026; Parkinson, 2022). Household survey data is commonly used by researchers studying the determinants of civilian attitudes and behaviour in post-conflict societies (e.g. Hartman et al., 2021; Voors et al., 2012; Fearon et al., 2015; Kao and Revkin, 2023; Peisakhin et al., 2025). Coverage error is an under-appreciated limitation on the external validity of survey findings in contexts affected by mass population movements. If not addressed, it can result in biased understandings of social and economic recovery after conflict, downplaying the role of forced displacement in shaping population outcomes.

[3]By residential accuracy we mean the percentage of visited points that correctly identified an inhabited residential building.

The remainder of this paper proceeds as follows: Section 2 describes the data sources; Section 3 details the methodology; Section 4 presents results; Section 5 investigates the sources of detection failure; and Section 6 discusses implications and concludes.

2 Data

2.1 Study Context and Community Selection

The 18 communities in our sample are located across the Anbar, Diyala, Ninewa and Salah al-Din governorates of northern Iraq, an area that experienced some of the most severe displacement associated with the Islamic State's territorial expansion between 2014 and 2017. At the height of the crisis, several million Iraqis were uprooted from these governorates, dispersed across IDP camps, urban areas, and informal settlements throughout the country. Following the military liberation of Mosul in 2017 and the gradual rollback of Islamic State control across the region, a substantial return movement began. By the time of our data collection in 2024–2025, large numbers of individuals had returned to their communities of origin, though a significant share of the displaced population remained in IDP camps or informally settled elsewhere. The result is a population structure that is simultaneously stationary and mobile: long-term residents, recently returned households, and IDP families may share the same street or building cluster, with no administrative record reliably distinguishing between them. Conventional sampling frame approaches — whether based on administrative population registers, in-person listing, or older satellite imagery — are poorly suited to this environment, since each is sensitive to the pace and direction of displacement and return and may therefore systematically fail to enumerate the groups of greatest policy interest.

The sampling frame described in this paper was constructed to support data collection of a household survey in Iraq. The surveys aimed to sample approximately 1,000 households drawn from the 18 communities. The goal was to characterise the local population in terms of economic and social well-being, in order to inform learning and evidence for programming delivered by the International Organization for Migration (IOM). The 18 communities were

drawn from IOM's operational caseload, selected on the basis of elevated IDP and returnee concentrations, economic and social vulnerability, and logistical accessibility for field teams. This purposive design was tailored to specific learning goals — typical of studies seeking to diagnose local needs or to evaluate the impact of an intervention — but the communities are not intended to be a representative sample of northern Iraq or Iraq nationwide. The patterns of ML-layer performance documented below may differ in communities with lower displacement rates, more stable building stock, or different conflict histories.

2.2 ML-Based Building Footprints

The primary input to our sampling frame is the GlobalML Building Footprints dataset (Microsoft, 2022), an open-source repository created by Microsoft in 2022 and made freely available under the Open Database License.[4] According to Microsoft's documentation, the dataset was constructed using convolutional neural networks trained on satellite imagery from Bing Maps, Maxar, Airbus, and IGN France; the model identifies building-like pixel patterns and converts them into georeferenced polygon outlines. For our study communities in northern Iraq, the available footprints were derived from imagery collected between 2014 and 2023. Each polygon centroid was extracted as a point, yielding the baseline candidate location for each detected structure.

2.3 Satellite Imagery

Freely available satellite basemap imagery was too outdated for validation purposes given the pace of construction, destruction, and resettlement in post-conflict northern Iraq. We therefore procured contemporaneous high-resolution imagery from Planet Labs using PlanetScope satellites, obtaining 8-band images at approximately 50cm ground resolution for each of the 18 communities. Images were corrected for surface reflectance and georeferenced to align with ground coordinates before being composited into community-level raster layers in QGIS. This up-to-date imagery layer served two functions: validating the ML-generated footprint points against current building states, and enabling the detection and

[4] At the time of writing, Iraq was not covered by Google's Open Buildings project, which provides comparable open-source footprints.

manual digitising of structures not present in the Microsoft dataset. To classify each community as urban or rural, we use the Global Human Settlement Population Layer (GHS-POP), a raster dataset produced by the European Commission's Joint Research Centre that combines satellite-derived built-up area estimates with census-based population data to generate gridded population density estimates at 100-metre resolution (Pesaresi et al., 2024). We use the 2015 vintage of GHS-POP — the edition most closely aligned with the pre-displacement baseline population — applying a threshold of 300 inhabitants per km^2 to distinguish Urban from Rural communities.

3 Methodology

3.1 Sampling Frame Construction

The sampling frame for each community was built through a three-step hybrid process. First, all algorithmically generated footprint points were validated against the Planet Labs imagery. Points marking locations that current satellite imagery confirmed did not contain a habitable structure — including demolished buildings, open ground, and vehicles — were flagged and removed from the candidate pool. Particular attention was paid to over-marking: the algorithm occasionally placed multiple centroids on a single structure, which would have inflated its probability of selection in a random draw. These duplicates were identified and reduced to a single representative point during validation. Figures 1 and 2 illustrate the resulting point layers for two communities in the sample; retained footprint points appear in green and deleted points in red, with clusters of red visible where over-marking or building demolition was detected.

Second, the Planet Labs imagery was used to identify structures absent from the footprint layer. These included newly constructed dwellings, recently reoccupied buildings, and informal shelters erected after the training imagery was collected. Analysts placed centroid points on each such structure in a dedicated QGIS layer, recording a timestamp and analyst identifier with each point. Screening criteria excluded partially collapsed structures judged uninhabitable, ancillary outbuildings immediately adjacent to a primary dwelling,

Figure 1: **8Shibat Community**

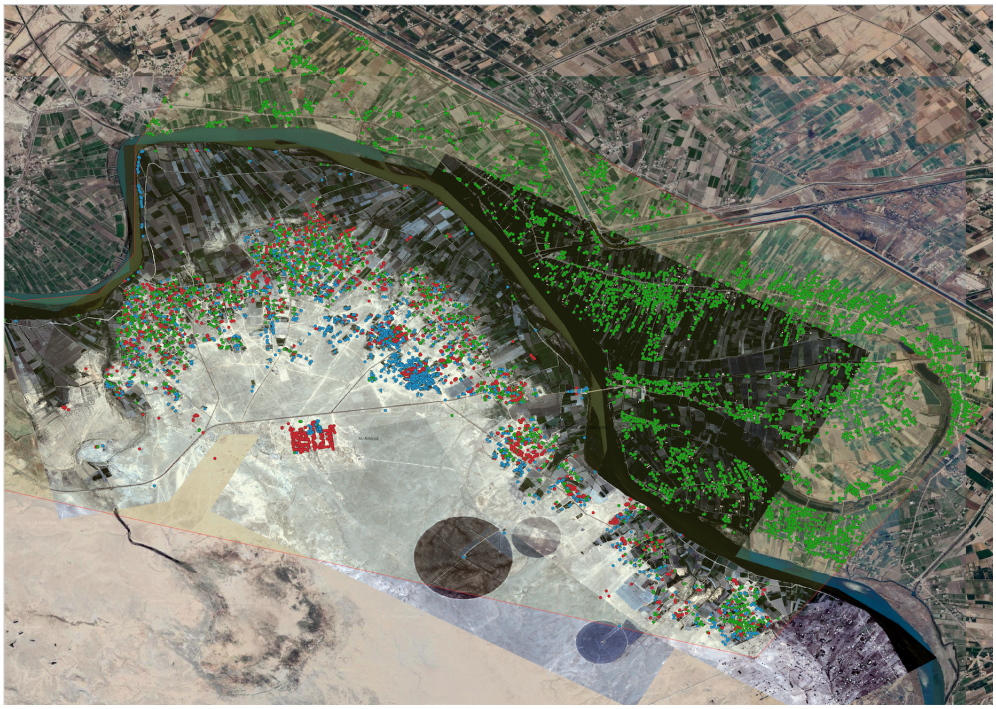

Figure 2: Bizaibiz Informal Settlement

and non-residential objects such as vehicles.

All digitising was performed by one of the authors, applying the same screening criteria uniformly across all 18 communities.

Single-analyst digitising is common in GIS validation work of this kind because it helps ensure that classification decisions are applied consistently across the dataset, without the need to reconcile potentially divergent judgments across multiple operators. At the same time, the choice between single and multiple coders involves trade-offs. Using several coders can provide a check on individual judgment and allow the assessment of inter-coder reliability, but it also requires procedures to harmonise differences in interpretation and may introduce additional variability in how decision rules are applied. In this application, the criteria governing inclusion and exclusion—habitability, residential use, and independence from an adjacent primary structure—are observable directly from high-resolution satellite imagery and admit relatively straightforward application; they do not require specialised domain knowledge beyond familiarity with the imagery and the coding rules. For this reason we rely on a single analyst and document the process carefully. The analyst identifier and timestamp recorded with each point allow the full digitising sequence to be audited if needed.

Manually placed points appear in blue in Figures 1 and 2; areas with dense blue coverage indicate neighbourhoods where recent development had substantially outpaced the footprint layer. Figure 3 provides a detailed view of the Bizaibiz informal settlement, where the value of current imagery is particularly visible. Similar maps for all other research sites are available in the Appendix.

Third, the validated footprint points and manually placed points were merged into a single sampling frame for each community, with point of origin — algorithmically derived or manually placed — recorded as an attribute. This attribute is the basis for the empirical comparisons in Section 4.

Figure 3: **Bizaibiz Informal Settlement Closeup**

3.2 Sampling Design

Our sampling frames constitute a proximity-weighted sample of the residential population within each community. Randomly selected geographic points were transmitted to IIACSS — a Baghdad-based survey firm and Iraq's sole GALLUP International affiliate (IIACSS, n.d.) — which conducted all field enumeration. Upon arrival at each selected point, enumerators recorded the state of the building and, where the structure was occupied and residential, proceeded with the survey. Points were visited in the order generated rather than reordered for logistical convenience, to avoid the geographic clustering bias associated with proximity-based selection (Lin and Kuwayama, 2016). To account for variation in local building density, all regression estimates reported in Section 4 use sampling weights proportional to each respondent's probability of selection. In the proximity-based design, a respondent's probability of selection is proportional to the number of eligible buildings within 300 metres of their interview point (`exposure.n`): respondents in denser areas, where more candidate structures are concentrated, are more likely to be reached. The weighted estimates therefore reflect the population distribution across the spatial density gradient of

the sampling frame.

3.3 Survey Administration

All fieldwork was conducted by IIACSS, whose enumerators administered the instrument in Arabic or Kurdish according to each respondent's preference. A typical interview lasted approximately 45–55 minutes. In addition to basic household characteristics — including gender composition, household size, and residential status (host, IDP, or returnee)[5] — the instrument asked a range of questions about economic and social conditions. Enumerators used handheld Android devices running the instrument in CAPI format; responses were transmitted electronically to the research team upon interview completion. The research team conducted systematic quality checks, reviewing recorded interview durations and flagging entries with implausibly short completion times or inconsistent responses for enumerator follow-up prior to analysis.

3.4 Estimation Strategy

To document the association between identification method and respondent composition, we estimate a separate exposure-weighted linear probability model (LPM) for each respondent type $k \in \{\text{host, IDP, returnee}\}$:

$$Y_{ik} \;=\; \alpha_k \;+\; \beta_k \cdot \text{Found}_i \;+\; \varepsilon_{ik}, \tag{1}$$

where $Y_{ik} = \mathbf{1}[\text{respondent } i \text{ belongs to type } k]$, $\text{Found}_i = \mathbf{1}[\text{point } i \text{ was placed manually via satellite image}$ and each observation is weighted by $w_i = \texttt{exposure.n}_i$, the number of eligible buildings within 300 metres of respondent i's interview point. Under weighted least squares, the intercept $\hat{\alpha}_k$ equals the weighted share of type k among ML-retained respondents; $\hat{\alpha}_k + \hat{\beta}_k$ equals the weighted share among manually found respondents; and the slope $\hat{\beta}_k$ is the exposure-weighted difference-in-means between the two groups, which is the estimand of interest.

[5]The survey instrument recorded residential status across four categories: host community member, IDP, returnee, and stayee — a household that remained in its community of origin throughout the conflict period without displacement. For the purposes of this analysis, stayees are grouped with host community members, as both represent established residents rather than displaced or returned populations.

Standard errors are heteroskedasticity-robust throughout. The model is estimated separately for each type k — rather than as a system — because the three outcomes sum to one and the OLS slope on Found_i is numerically identical across specifications.

The exposure weight w_i corrects for the proximity-based nature of the sampling design. In a setting where survey points are drawn from a frame of building centroids, a respondent located in a high-density area is more likely to be reached because more candidate points fall within proximity of their dwelling. Weighting by $\texttt{exposure.n}_i$ — the local count of eligible buildings — downweights respondents in dense areas and upweights those in sparse areas, so that the weighted estimators reflect the distribution of the full building population rather than the sample of visited points. Appendix Table A4 confirms that the substantive conclusions are robust to replacing these weighted estimates with unweighted OLS.

4 Results

When creating our sampling frame, our team examined a combined area of 210.20 km^2 across 18 communities selected for their high IDP and returnee populations, levels of economic and social vulnerability, and field accessibility. Table A5 in the appendix provides the full community-level breakdown — including individual site areas, algorithmic and manual point counts, deletion rates, and residential accuracy — for each of the 18 communities. Our communities initially contained 52,448 buildings identified by Microsoft's machine learning repository, supplemented by 9,660 buildings identified manually through satellite imagery review. During validation, 505 algorithmically generated points were marked for deletion, yielding a combined total of 61,603 viable points for the survey teams. Deleted points can be seen in Figure 3, which shows red points marking locations that satellite imagery confirmed did not contain habitable structures. These figures are summarised in Table 1; the note to that table also reports the total analyst time required to digitise the found points.

To measure the labour cost of the manual correction step, we recorded a timestamp for each manually placed point in QGIS, and treated consecutive gaps of more than 20 minutes as analyst breaks rather than active digitising time. This approach ensures the reported

Table 1: Building Sampling Frame Summary

Total Area	210.20 km^2
Starting Points	52,448
Deleted Points	505
Found Points	9,660
Ending Points	61,603

Note: Reports the distribution of each type of geospatial point marked in the region. Data is sourced from the datasets containing additional attributes for found points, as well as datasets containing the extent of each community and new building information. Manually digitising the 9,660 found points required approximately 118.4 analyst-hours (0.56 hours per km^2), excluding breaks defined as gaps of more than 20 minutes between consecutive clicks.

figure reflects time spent actively clicking points and validating algorithmic output rather than including extended idle periods. The resulting time estimate — reported in the note to Table 1 — provides a direct basis for practitioners evaluating whether this methodology is feasible given their available resources, and represents one of the few published estimates of the per-area cost of hybrid sampling frame construction in a conflict setting.

Of the 61,603 viable points, 1,225 were visited by the survey team. Of these, 23 were commercial buildings, 46 were construction sites, 23 were destroyed, 31 had no building present, 41 were abandoned, and 1,061 had residents who completed the survey. As shown in Table 2, approximately 87% of visited points were occupied residential buildings, confirming a high capture rate for actual domiciles. This result validates the use of satellite imagery to pre-screen building state prior to field deployment, reducing enumerator time spent visiting uninhabitable structures. Tables A2 and A3 in the appendix further disaggregate building status by community type (urban versus rural) and by identification method (original/kept versus found), respectively, for readers interested in whether building condition differs systematically across these dimensions.

Among the 1,061 respondents who completed the survey — 823 reached via algorithmically retained points and 238 via manually placed found points — we can classify each by whether the point that led to their interview was an algorithmically generated footprint retained during validation (Original and Kept) or a point placed manually using up-to-date satellite imagery (Found). This distinction is central to the paper's measurement contribu-

Table 2: State of Buildings Visited

Building State	N	%	% of Buildings in Mapped Area
Yes, residential building (including informal or temporary shelter)	1061	86.61	91.80
No, commercial, industrial or administrative building	23	1.88	1.01
No, construction site (nobody living inside)	46	3.76	2.21
No, destroyed building	23	1.88	1.16
No, there is no building	31	2.53	1.74
No, abandoned house	41	3.35	2.10
Total	*1225*	*NA*	*NA*

Note: Table reports the condition of buildings visited by each survey team. Data is sourced from the survey dataset recorded by each survey team.

tion. Whether the composition difference between identification methods reflects properties of the sampling approach itself or the geographic co-location of building types and population groups, the gap directly quantifies the coverage error a researcher would incur by forgoing the manual correction step. Table 3 documents this comparison by estimating a separate exposure-weighted linear probability model for each respondent type — host, returnee, and IDP — regressing group membership on an indicator for manual identification. The intercept recovers the weighted share of each group among original and kept points; the sum of intercept and slope recovers the share among found points; and the slope, reported in the Difference column, gives the OLS-estimated gap, with stars indicating the p-value of the associated t-test.

The results show that identification method is meaningfully associated with respondent composition. Hosts are substantially underrepresented among found points relative to the algorithmic baseline, a difference that is statistically significant, consistent with hosts being more likely to occupy well-established structures already captured by the algorithm's training data. Returnees are found at a statistically significantly higher rate among manually placed points, suggesting that the more recent satellite imagery used in the manual correction step was better able to locate the structures in which returnees tend to reside — newly constructed dwellings and recently reoccupied buildings that the footprint layer, trained on older imagery, had not yet recorded. The IDP share is broadly comparable across methods, with a small

and weakly significant difference. Table A4 in the appendix replicates this analysis without exposure weights and confirms that the pattern of results is robust to the weighting choice.

Table 3: Share of Respondent Types by Identification Method

	Original and Kept Points (%)	Found Points (%)	Difference (pp)
Host	23.01	12.20	-10.81***
Returnee	59.14	62.92	3.78
IDP	17.84	24.87	7.03**

Note: Table presents share of respondent types by identification method. *p<0.10, **p<0.05, ***p<0.01. Difference column reports the OLS coefficient on an indicator for manual identification (Found = 1, Original/Kept = 0), estimated separately for each respondent type. Positive values indicate a higher share among found points. Estimates use exposure-weighted OLS; see Table A3 for a comparison with unweighted estimates.

The aggregate results in Table 3 suggest that the value of manual correction is not uniform across the sample. To investigate this, Table 4 disaggregates the analysis by community settlement type, using population density estimates from the Global Human Settlement Layer (GHS-POP) to classify each community as either Urban ($\geq$300 inhabitants per km^2, encompassing the GHS Urban Cluster and Urban Center categories) or Rural ($<$300 inhabitants per km^2). Six communities are classified as Rural and twelve as Urban. Within each settlement type, we estimate the same set of weighted linear probability models as in Table 3, reporting the share of each respondent group among original/kept and found points together with the OLS-estimated difference.

The results reveal a sharp urban–rural gradient. In rural communities, the differences in respondent composition between found and original/kept points are larger and, for IDPs and returnees, more precisely estimated than in the pooled analysis.This is consistent with rural settings presenting greater challenges for ML-based building detection: lower building density, less standardised rooftop materials, and a higher prevalence of informal and temporary structures reduce the coverage of training-data-derived footprints, making the manual correction step particularly valuable. In urban communities, where the ML layer performs more reliably and building stock is more stable, the differences between found and original/kept points are smaller and less consistently significant. This pattern implies that practitioners working in predominantly urban environments may be able to place greater reliance on the

ML layer alone, while those working in mixed or rural settings should plan for more intensive manual review.

Table 4: Share of Respondent Types by Community Status

	Original and Kept Points (%)	Found Points (%)	Difference (pp)
Rural			
Host	20.68	13.43	-7.25
IDP	21.20	31.67	10.46**
Returnee	58.12	54.91	-3.21
Urban			
Host	24.02	11.36	-12.66***
IDP	16.16	20.19	4.04
Returnee	59.82	68.45	8.62*

Note: Table presents share of respondent types by community status. *p<0.10, **p<0.05, ***p<0.01. Difference column reports the OLS coefficient on an indicator for manual identification (Found = 1, Original/Kept = 0), estimated separately for each respondent type within each community type. Positive values indicate a higher share among found points.

The results in Tables 3 and 4 document differences in composition *between* the two identification methods. A natural follow-on question is what those differences imply for the aggregate composition of the survey sample: if the manual correction step had been omitted entirely and the researcher had deployed the footprint layer alone, how much would the estimated population shares have differed from those obtained under the hybrid approach? Table 5 answers this directly. For each respondent type, it compares the weighted share under a footprint-only frame — restricted to respondents reached via original or kept algorithmically generated points — with the weighted share under the actual hybrid frame, which includes both retained and manually found respondents. The difference, reported in the Bias column, is the measurement error in population-share terms that a researcher would incur by forgoing manual validation.

The results confirm that the footprint layer alone would produce a systematically biased picture of the community's population composition. Hosts are overrepresented among retained points relative to the full hybrid sample, consistent with hosts disproportionately occupying established structures that the training data reliably captured. IDPs and returnees are underrepresented under the footprint-only frame, as both groups are more likely

Table 5: Counterfactual Frame Bias: Footprint-Only vs. Hybrid Composition

	Footprint-Only Frame (%)	Hybrid Frame (%)	Bias (pp)
Overall			
Host	23.01	20.22	2.79
IDP	17.84	19.66	-1.82
Returnee	59.14	60.12	-0.98
Rural			
Host	20.68	18.48	2.20
IDP	21.20	24.37	-3.17
Returnee	58.12	57.15	0.97
Urban			
Host	24.02	21.04	2.98
IDP	16.16	17.11	-0.95
Returnee	59.82	61.85	-2.03

Note: Footprint-Only Frame reports the exposure-weighted share of each respondent type among respondents reached via original or kept algorithmically generated points only. Hybrid Frame reports the weighted share across all respondents (retained and manually found). Bias = Footprint-Only − Hybrid: positive values indicate overestimation of that group under a footprint-only approach. Urban/Rural classification from GHS-POP ($\geq$300 inhabitants/km^2 = Urban).

to reside in the newer or informally constructed structures that only the current satellite imagery could identify. The bias is concentrated in rural communities, where the coverage limitations of the footprint layer are most severe. In urban communities, where building stock is more stable and the algorithm performs reliably, the footprint-only and hybrid compositions are broadly similar and the implied bias is small. This decomposition sharpens the operational guidance suggested by Table 4: for practitioners deciding how intensively to conduct manual validation, the rural bias estimates in Table 5 provide a direct, quantitative basis for that prioritisation decision. If the goal is to enumerate IDPs and returnees accurately — the populations of primary interest in most humanitarian survey contexts — the cost of skipping manual correction is not uniform across geography, and is largest precisely where these groups are most concentrated. It bears noting that the counterfactual in Table 5 is constructed by restricting to respondents reached via retained points from the validated frame — that is, points that survived the deletion step. A researcher who deployed the raw, unvalidated footprint layer would additionally have dispatched survey teams to the 505 over-marked and demolished-structure points. Because those points mark non-existent or

Figure 4: **Closeup of Bizaibiz Informal Settlement**
Left: Without up-to-date SkySat Imagery. Right: Including recent SkySat imagery

duplicated structures, the visits would yield no completed surveys and therefore would not alter the estimated respondent composition. The composition bias estimates in Table 5 thus accurately characterise what a footprint-only frame would produce. What those estimates do not capture is the additional field cost: those 505 wasted visits represent real operational expenditure, meaning that Table 5 understates the full programmatic cost of forgoing manual validation.

5 Sources of Detection Failure

The urban–rural heterogeneity documented in Table 4 suggests that specific characteristics of the building stock — rather than the algorithm's general accuracy — drive the coverage gap between retained and manually identified points. Two candidate mechanisms are apparent. First, temporary and informal shelters are not detected by building footprint algorithms, which are trained on satellite imagery of permanent roof structures; if displaced populations disproportionately occupy such shelters, the footprint layer will systematically miss them. Second, wartime aerial bombardment and artillery fire destroy or alter buildings in ways that degrade algorithmic detection, while the reconstruction that follows creates new structures visible in current satellite imagery but absent from the training data. We test each mechanism in turn.

5.1 Shelter Informality and Algorithmic Invisibility

Table 6 reports exposure-weighted linear probability models regressing the Found indicator on measures of shelter informality. Respondents living in informal shelters — tents, makeshift structures, or prefabricated caravans — are 20 percentage points more likely to have been identified via manual satellite review than via the footprint layer ($p = 0.001$). Restricting to tent-dwellers, the estimate rises to 34 percentage points ($p < 0.001$). Among tent-dwellers who were interviewed, roughly half were reached at manually placed points — a group that would have been entirely absent from a footprint-only sampling frame.

Table 6: Shelter Type and Algorithmic Detection Failure

	Coefficient	Std. Error	N
(1)			
Constant (permanent shelter)	0.245***	(0.013)	
Informal shelter (tent/makeshift/prefab)	0.202***	(0.062)	1106
(2)			
Constant (non-tent)	0.247***	(0.014)	
Tent only	0.341***	(0.075)	993

Note: Dependent variable: Found (= 1 if building was manually identified via satellite imagery, 0 if algorithmically retained). Each panel reports a separate exposure-weighted LPM. The constant gives the Found rate among the omitted reference group: respondents in permanent shelter (panel 1) and non-tent respondents (panel 2). Informal shelter includes tent, makeshift shelter, and prefab/caravan/RHU. *p<0.10, **p<0.05, ***p<0.01.

The shelter mechanism predicts that controlling for shelter type should attenuate the IDP–Found association documented in Table 3, since IDPs are disproportionately concentrated in informal shelters. This concentration is itself highly heterogeneous: 48% of IDPs in rural communities live in tents, compared with zero among urban IDPs. Table 7 confirms the prediction. In rural communities, controlling for informal shelter reduces the IDP coefficient by 107% — more than fully explaining the rural IDP–Found gap. In urban communities, where IDPs do not live in tents, the attenuation is 1.2%. The shelter mechanism is a rural phenomenon: in settings where displaced populations occupy tents and makeshift structures that the algorithm cannot detect, manual satellite review is the only way to enumerate them.

The Bizaibiz informal settlement (Figure 3) provides the clearest illustration. Eighty-

three percent of respondents are IDPs, 79% live in tents, 54% were reached at manually placed points, and the community experienced no recorded wartime bombardment. The dense cluster of blue points in Figure 2 marks the locations where the footprint layer failed entirely — a settlement that, for the purposes of a footprint-only frame, does not exist.

Table 7: Mediation: Does Shelter Type Explain the IDP–Found Gap?

Sample	IDP (baseline)	IDP (+ shelter control)	Attenuation
All communities	0.085**	0.058	32.4%
Rural	0.12**	-0.008	106.6%
Urban	0.051	0.051	1.2%

Note: Dependent variable: Found. Baseline column reports the coefficient on an IDP indicator from an exposure-weighted LPM without controls. Shelter control column adds an indicator for informal shelter (tent, makeshift, or prefab). Attenuation = percentage reduction in the IDP coefficient after adding the shelter control. Values above 100% indicate the IDP coefficient changes sign. *p<0.10, **p<0.05, ***p<0.01.

5.2 Wartime Bombardment and Post-Conflict Reconstruction

The shelter mechanism explains why IDPs are over-represented at manually placed points but does not account for the returnee patterns in Tables 3 and 4. Returnees generally occupy permanent structures and should therefore appear in the footprint layer at rates comparable to hosts. That they do not suggests a second channel. In communities that experienced heavy aerial bombardment during the 2014–2017 anti-ISIS campaign, destruction and subsequent reconstruction reshaped the building stock. The training imagery, derived from pre-conflict or early-conflict satellite passes, could not have recorded buildings that were destroyed and rebuilt in the intervening years. Returnees resettling in these rebuilt homes would therefore be systematically located at manually identified points.

To classify communities by wartime conflict intensity, we use the UCDP Georeferenced Event Dataset (GED v25.1; Davies et al., 2024), which records individually geolocated conflict events in Iraq from 1989 onward. We restrict to the period of the anti-ISIS campaign (2014–2017) and classify events involving aerial bombardment, shelling, artillery, rocket, and mortar attacks via pattern matching on source descriptions.[6] Events falling within a 5km

[6]The UCDP GED does not record a structured event subtype comparable to ACLED's `sub_event_type`. We classify events by searching the `source_headline` and `source_article` fields for terms indicating aerial

buffer of each community centroid are counted and normalized by community area. Because only 18 communities yield a limited set of distinct density values, we divide communities into terciles of bombardment intensity by rank rather than by continuous quantile cutpoints.

Table 8 replicates the respondent-composition analysis of Table 3 separately within each bombardment tercile. The results reveal a pattern that aggregate estimates conceal. In low-bombardment communities, the composition gap follows the shelter channel: IDPs are over-represented at manually placed points by 27 percentage points ($p < 0.001$), while returnees are under-represented by 16 percentage points ($p = 0.03$). In high-bombardment communities, the pattern reverses: returnees are over-represented at manually placed points by 17 percentage points ($p = 0.03$), while IDPs show no significant difference. The reversal is consistent with the reconstruction mechanism — in communities where aerial bombardment destroyed residential structures, post-conflict rebuilding generated new buildings visible in current satellite imagery but absent from the training data. The aggregate returnee coefficient in Table 3, which pools across bombardment intensity, masks this heterogeneity: returnees are under-represented at Found points in communities where little reconstruction occurred, and over-represented where reconstruction was extensive.

Table 9 presents a combined specification that enters both mechanisms simultaneously. Column (1) includes respondent type and settlement type; columns (2) and (3) add informal shelter and self-reported house destruction; columns (4) and (5) add the ACLED and UCDP bombardment terciles, respectively, as alternative codings of conflict intensity. In the full specification, informal shelter remains strongly associated with manual identification (+30 percentage points, $p < 0.001$ in column (5)), while self-reported house destruction is not. Adding the bombardment tercile substantially raises the R^2 relative to column (3): from 0.028 to 0.071 under the ACLED coding and to 0.095 under the UCDP coding. The bombardment-tercile coefficients in the combined model are smaller than the split-sample

or artillery attack: strong terms (airstrike, artillery, howitzer, shelling, air raid, precision strike) are counted regardless of context, while weak terms (bomb, rocket, mortar, missile) are counted only when no IED, car-bomb, or suicide-attack compound (*car bomb*, *suicide bomber*, VBIED, *improvised explosive*, etc.) is present in the source text. IED and car-bomb events are thereby excluded by construction, mirroring ACLED's structured separation of aerial/artillery events from remote-explosive/IED events. Of 2,818 Iraq events during 2014–2017, 311 (11.0%) are classified as aerial or artillery.

Table 8: Respondent Composition by Identification Method, Split by UCDP Bombardment Intensity (2014–2017)

	Original/Kept (%)	Found (%)	Difference (pp)
Low bombardment			
Host	24.29	12.72	-11.58**
Returnee	56.09	40.55	-15.53**
IDP	19.62	46.73	27.11***
Medium bombardment			
Host	11.76	9.65	-2.12
Returnee	80.84	74.14	-6.69
IDP	7.40	16.21	8.81***
High bombardment			
Host	41.12	28.33	-12.79
Returnee	25.10	41.91	16.82**
IDP	33.78	29.75	-4.02

Note: Replicates Table 4 within terciles of UCDP *aerial and artillery* bombardment intensity. Event density is computed as regex-classified aerial/artillery events (airstrike, artillery, howitzer, shelling, air raid, or bomb/rocket/mortar/missile when no IED/car-bomb/suicide-attack compound is present) per km^2 within a 5km buffer of the community centroid, 2014–2017. Communities are ranked by aerial/artillery events per km^2 and split at tercile boundaries. IED and car-bomb events are excluded by construction. Difference column reports the exposure-weighted LPM coefficient on the Found indicator, estimated separately for each respondent type within each tercile. *p<0.10, **p<0.05, ***p<0.01.

patterns documented in Tables 8 and A6, consistent with bombardment intensity being partially absorbed by the urban indicator once both enter the regression jointly.[7]

These results identify two distinct sources of detection failure that operate on different populations in different settings. Shelter informality causes the footprint layer to miss IDPs in rural areas, where temporary and improvised structures lack the rooftop signatures on which building detection algorithms are trained. Wartime bombardment causes the layer to miss returnees in heavily damaged communities, where post-conflict reconstruction has reshaped the building stock since the training imagery was collected. The opposing directions of the returnee coefficient across low- and high-bombardment communities — negative in

[7]As a robustness check, we replicate the bombardment analysis using ACLED conflict data (2016–2023), which records structured event subtypes but begins two years after the onset of major anti-ISIS operations. The qualitative pattern — IDPs over-represented at Found points in low-bombardment communities, returnees over-represented in high-bombardment communities — is unchanged. These results appear in Appendix Tables A6 and A7.

Table 9: Determinants of Manual Identification (Found)

	(1)	(2)	(3)	(4)	(5)
IDP	0.164***	0.142***	0.156***	0.149***	0.140***
	(0.044)	(0.045)	(0.046)	(0.045)	(0.044)
Returnee	0.113***	0.118***	0.120***	0.046	0.018
	(0.036)	(0.035)	(0.036)	(0.037)	(0.037)
Informal shelter		0.161**	0.186***	0.169**	0.301***
		(0.067)	(0.070)	(0.069)	(0.069)
House still destroyed			-0.054	-0.027	-0.015
			(0.040)	(0.039)	(0.039)
ACLED bombardment: Medium				0.130***	
				(0.035)	
ACLED bombardment: High				-0.123***	
				(0.041)	
UCDP bombardment: Medium					0.262***
					(0.034)
UCDP bombardment: High					-0.052
					(0.041)
Urban	-0.061**	-0.051*	-0.053*	-0.077**	0.081**
	(0.029)	(0.029)	(0.029)	(0.032)	(0.033)
Intercept	0.198***	0.185***	0.190***	0.227***	0.044
	(0.037)	(0.037)	(0.037)	(0.040)	(0.042)
N	991	991	982	982	982
R^2	0.021	0.026	0.028	0.071	0.095

Note: Dependent variable: Found (= 1 if building was manually identified via satellite imagery, 0 if algorithmically retained). All specifications use exposure-weighted LPM. Reference categories: Host (respondent type), Rural (settlement type), house not destroyed (destruction status), Low bombardment (tercile). Column (1): respondent status and settlement type only. Column (2): adds informal shelter indicator. Column (3): adds house destruction indicator. Column (4): adds ACLED bombardment tercile (communities ranked by aerial/artillery events per km^2 within 5km buffer, 2016–2023, split at tercile boundaries). Column (5): adds UCDP bombardment tercile (regex-classified aerial/artillery events per km^2, 2014–2017). *p<0.10, **p<0.05, ***p<0.01.

the former, positive in the latter — explain why the pooled estimate in Table 3 understates the heterogeneity of coverage error across settings. For practitioners, the implication is that the value of manual validation depends not only on whether a community is urban or rural (Table 4) but also on its conflict history and the prevalence of informal shelter — information that is often available before a sampling frame is constructed.

6 Conclusion

GIS-based sampling frames have long been recognized as advantageous in conflict-affected environments precisely because satellite imagery enables safe, unbiased enumeration without requiring field presence (Lin and Kuwayama, 2016). This paper extends that tradition by integrating algorithmically generated building footprints into the workflow and providing, to our knowledge, the first systematic evidence in a social science context on the coverage implications of algorithmically generated building footprints for population measurement in a post-conflict setting.

Our residential accuracy rate of 87% confirms that the hybrid workflow produces a sampling frame with strong coverage of actual domiciles even in a post-conflict environment with significant building churn. Fewer than 1% of algorithmically generated points required deletion during validation, and manually placed points account for approximately 15% of the total — indicating that the footprint layer provides a reliable and comprehensive starting point that requires targeted supplementation rather than wholesale replacement. Among the 505 deleted points, the most common reason was over-marking: the algorithm placed multiple points on single structures, which would have inflated their probability of selection in a random draw and biased empirical results. This clustering is visible in Figure 3, where deletion points concentrate on individual buildings in densely built areas. Without validation, the probability of such buildings being selected would have been elevated in proportion to the number of duplicate points, introducing a systematic bias into the randomisation.

The analysis in Tables 3 and 4 demonstrates that identification method is systematically associated with respondent composition in ways that are substantively meaningful for survey design and inference, and Table 5 translates those differences into a direct estimate of the bias a footprint-only frame would have introduced. Manually placed points are associated with a significantly higher share of returnees and a significantly lower share of hosts relative to retained points, while the IDP share is broadly comparable across methods at the aggregate level. These patterns are consistent with a structural interpretation of the two approaches: the footprint layer, trained on imagery that predates recent conflict and displacement, is

better calibrated to established residential structures where hosts are more likely to reside, while the manual correction step — using current satellite imagery — more readily identifies recently occupied or newly constructed structures. Section 5 tests this interpretation directly, identifying two distinct mechanisms — shelter informality and post-bombardment reconstruction — that explain the composition gaps and their heterogeneity across settings. The importance of this distinction extends beyond descriptive analysis: if the frame is used to weight survey estimates back to a population total, the choice of identification method affects which population is effectively being targeted.

The heterogeneity documented in Tables 4 and 8 sharpens these conclusions considerably. The gap between retained and manually found respondent compositions is concentrated in rural communities and varies systematically with conflict intensity. The mechanism analysis in Section 5 identifies two sources: shelter informality, which fully explains the rural IDP–Found gap, and post-bombardment reconstruction, which drives the returnee–Found association in heavily damaged communities. These are not transient artifacts of active conflict — our data were collected roughly a decade after the bulk of the fighting, and the detection biases persist, indicating that the algorithm's calibration to pre-conflict building stock remains a durable source of coverage error long after hostilities end. For practitioners, the implication extends beyond the simple urban–rural gradient: the effort required for intensive manual validation should be concentrated in rural communities with high displacement prevalence and in communities with histories of heavy wartime destruction, where reconstruction has most substantially reshaped the building stock since the training data was collected.

This paper documents a methodology for sampling frame construction in post-conflict environments and provides, to our knowledge, the first systematic evidence in a social science context on how algorithmically generated and manually placed sampling frame points differ in respondent composition. Several limitations bear noting. The 18 communities studied were selected on the basis of high IDP and returnee populations, economic vulnerability, and field accessibility — a purposive design appropriate for the substantive goals of the survey research but one that limits the representativeness of the sample. Communities with

lower displacement rates, more stable building stock, or different conflict histories may exhibit different patterns of footprint-layer performance. A further limitation is the absence of an external benchmark for the true population composition of the study communities. The paper documents that retained and manually placed points are associated with different respondent compositions, and the counterfactual in Table 5 quantifies the gap between them. The implicit assumption — that the hybrid frame is closer to the actual population distribution than the footprint-only frame — is plausible given that manual validation uses current satellite imagery while the footprint layer was trained on pre-conflict data, but it is not directly testable without an independent enumeration against which to validate either frame.

The approach relies entirely on open-source tools — QGIS and the Microsoft Footprints repository — and commercially accessible satellite imagery. The primary labour cost is the manual validation and digitising step: across our 18 communities spanning 210.20 km^2, this required 118.44 analyst-hours, equivalent to approximately 0.56 hours per km^2 or 6.6 hours per community on average. At the point level, each manually placed building required roughly 44 seconds of active digitising time. These figures are directly transferable to other operations: practitioners can multiply the hours-per-km^2 rate by their planned area of coverage to project total analyst time before committing to field deployment, without needing to conduct a pilot or assume comparability with in-person listing costs. Beyond the Iraqi context, the methodology is applicable wherever accurate population measurement is needed in environments where traditional enumeration is impractical. While most relevant for population monitoring in displacement contexts, the approach has applications for demographic measurement, humanitarian policy delivery, and survey sampling across a broader set of highly dynamic demographic settings — including areas undergoing conflict-driven destruction and reconstruction, high-mobility contexts where populations occupy tents or temporarily vacant structures, informal and squatter settlements, and post-disaster environments where building stock changes rapidly. For practitioners working in such settings, the urban–rural heterogeneity documented here supports a tiered operational strategy — more intensive manual review in rural and mixed-density communities, where Table 5 shows the

footprint-only bias to be largest and the most vulnerable populations most systematically undercounted, and lighter validation in stable urban areas where the algorithm performs reliably and the implied bias is small. The numbers reported here support a quantitative decision rule for allocating manual validation effort. In our data, the manual digitising step required 0.56 analyst-hours per km^2. Pairing this cost with the bias estimates in Table 5 allows a practitioner to estimate the return on that investment in terms of coverage-error reduction. Consider, for example, a rural community of 10 km^2: manual validation would require approximately 5.6 analyst-hours and, based on our estimates, would reduce the IDP undercount by roughly 3.2 percentage points and the host overcount by roughly 2.2 percentage points relative to a footprint-only frame. In an urban community of the same size, the same 5.6 hours would correct an IDP undercount of less than 1 percentage point. The trade-off is stark: in rural settings, each analyst-hour invested yields a substantially larger reduction in coverage error than in urban ones. A practitioner facing a fixed budget of analyst time should therefore prioritise rural and peri-urban communities — where the footprint layer's coverage gaps are largest and the populations of greatest policy interest are most systematically undercounted — and allocate lighter validation to stable urban areas where the implied bias is small. These cost-per-percentage-point figures are calibrated to the northern Iraq context studied here; practitioners in other settings should treat them as illustrative benchmarks to be updated with local validation data where available.

References

Admasu, Y, S Alkire, UE Ekhator-Mobayode, F Kovesdi, J Santamaria, and S Scharlin-Pettee, "A multi-country analysis of multidimensional poverty in contexts of forced displacement," 2021.

Aguilera, A., N. Krishnan, D. Sharma, and T. Vishwanath, "Sampling for Representative Surveys of Displaced Populations," in "Forced Displacement and the Developing World," Cham: Palgrave Macmillan, 2019.

Aiken, E., S. Bellue, D. Karlan, C. Udry, and J. Blumenstock, "Machine learning and phone data can improve targeting of humanitarian aid," *Nature*, 2022.

Alrababah, Ala, Marine Casalis, Daniel Masterson, Dominik Hangartner, Stefan Wehrli, and Jeremy Weinstein, "Reducing attrition in phone-based panel surveys: best practices and semi-automation for survey workflows," *Political Science Research and Methods*, 2026, *14* (1), 221–230.

Bauer, J., "Selection Errors of Random Route Samples," *Sociological Methods and Research*, 2014.

Boo, G., E. Darin, D. R. Leasure, C. A. Dooley, H. R. Chamberlain, A. N. Lazar, and A. J. Tatem, "High-resolution population estimation using household survey data and building footprints," *Nature Communications*, 2022, *13*, 1330.

Bowles, Jeremy, "Identifying the Rich: Registration, Taxation, and Access to the State in Tanzania," *American Political Science Review*, 2024, *118* (2), 602–618.

Davies, Shawn, Therese Pettersson, and Magnus Öberg, "Organized Violence 1989–2023, and the Return of Conflict between States," *Journal of Peace Research*, 2024, *61* (4).

Escamilla, V., M. Emch, L. Dandalo, W. C. Miller, F. Martinson, and I. Hoffman, "Sampling at community level by using satellite imagery and geographical analysis," *Bulletin of the World Health Organization*, 2014.

Fearon, James D., Macartan Humphreys, and Jeremy M. Weinstein, "How Does Development Assistance Affect Collective Action Capacity? Results from a Field Experiment in Post-Conflict Liberia," *American Political Science Review*, 2015, *109* (3), 450–469.

Fransen, S. and O. Bilgili, "Who reintegrates? The constituents of reintegration of displaced populations," *Population, Space, and Place*, 2018.

_ , I. Ruiz, and C. Vargas-Silva, "Return Migration and Economic Outcomes in the Conflict Context," *World Development*, 2017.

Galway, L., N. Bell, A. Hagopian, S. Al Shatari, G. Burnham, A. Flaxman, W. Weiss, J. Rajaratnam, and T. Takaro, "A two-stage cluster sampling method

using gridded population data, a GIS, and Google Earth^{TM} imagery in a population-based mortality survey in Iraq," *International Journal of Health Geographics*, 2012.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau, *Survey Methodology*, 2 ed., Hoboken, NJ: Wiley, 2009.

Hartman, Alexandra C, Benjamin S Morse, and Sigrid Weber, "Violence, displacement, and support for internally displaced persons: Evidence from Syria," *Journal of Conflict Resolution*, 2021, *65* (10), 1791–1819.

IIACSS, "IIACSS Website," https://iiacss.org/. Accessed: 2025-02-16.

International Organization for Migration, "DTM South Sudan - Urban Multi-Sector Needs, Vulnerabilities and COVID 19 Impact Survey (FSNMS+) - Juba Town," *IOM*, 2021.

Kamedjeu, R., "Tracking the polio virus down the Congo River: a case study on the use of Google Earth^{TM} in public health planning and mapping," *International Journal of Health Geographics*, 2009.

Kao, Kristen and Mara R. Revkin, "Retribution or Reconciliation? Post-Conflict Attitudes toward Enemy Collaborators," *American Journal of Political Science*, 2023, *67* (2), 358–373.

Kondylis, F., "Conflict displacement and labor market outcomes in post-war Bosnia and Herzegovina," *Journal of Development Economics*, 2010.

Landry, Pierre F. and Mingming Shen, "Reaching Migrants in Survey Research: The Use of the Global Positioning System to Reduce Coverage Bias in China," *Political Analysis*, 2005, *13* (1), 1–22.

Lee, Melissa M. and Nan Zhang, "Legibility and the Informational Foundations of State Capacity," *The Journal of Politics*, 2017, *79* (1), 118–132.

Lin, Y. and D. Kuwayama, "Using satellite imagery and GPS technology to create random sampling frames in high risk environments," *International Journal of Surgery*, 2016.

Microsoft, "GlobalMLBuildingFootprints," https://github.com/microsoft/GlobalMLBuildingFootprints 2022. Accessed: 2025-01-22.

Neal, I., S. Seth, G. Watmough, and M. S. Diallo, "Census-independent population estimation using representation learning," *Scientific Reports*, 2022, *12*, 5185.

O'Reilly, Colin, "Household Recovery from Internal Displacement in Northern Uganda," *World Development*, 2015, *76*, 203–215.

Parkinson, Sarah E., "(Dis)courtesy Bias: "Methodological Cognates," Data Validity, and Ethics in Violence-Adjacent Research," *Comparative Political Studies*, 2022, *55* (3), 420–450.

Peisakhin, Leonid, Nik Stoop, and Peter Van Der Windt, "Who Hosts? The Correlates of Hosting the Internally Displaced," *American Political Science Review*, 2025, *119* (3), 1143–1158.

Pesaresi, M., M. Schiavina, P. Politis, S. Freire, K. Krasnodebska, J. H. Uhl, and T. Kemper, "Advances on the Global Human Settlement Layer by joint assessment of Earth Observation and population survey data," *International Journal of Digital Earth*, 2024, *17* (1).

Ruiz, Isabel and Carlos Vargas-Silva, "The Legacies of Armed Conflict: Insights From Stayees and Returning Forced Migrants," *Journal of Conflict Resolution*, 2025, *69* (1), 17–45.

Scott, James C., *Seeing like a state : how certain schemes to improve the human condition have failed / James C. Scott.* Veritas Paperbacks Ser., New Haven: Yale University Press, 1998.

Shaver, Andrew, Benjamin Krick, Judy Blancaflor, Xavier Liu, Ghassan Samara, Sarah Yein Ku, Shengkuo Hu, Joshua Angelo, Martha Carreon, Trishia Lim, and et al., "The Causes and Consequences of Refugee Flows: A Contemporary Reanalysis," *American Political Science Review*, 2025, *119* (1), 526–534.

Steele, J., P. Sundsoy, C. Pezzulo, V. Alegana, T. Bird, J. Blumenstock, J. Bjelland, K. Engo-Monsen, Y. de Montoye, A. Iqbal, H. Hadiuzzaman, X. Lu, E. Wetter, A. Tatem, and L. Bengtsson, "Mapping poverty using mobile phone and satellite data," *Journal of the Royal Society Interface*, 2017.

Tellez, Juan F. and Laia Balcells, "Social Cohesion, Economic Security, and Forced displacement in the Long-run: Evidence From Rural Colombia," *Journal of Conflict Resolution*, 2025, *69* (1), 46–73.

Thomson, D. R., D. A. Rhoda, A. J. Tatem, and M. C. Castro, "Gridded population survey sampling: a systematic scoping review of the field and strategic research agenda," *International Journal of Health Geographics*, 2020, *19*, 34.

_ , F. R. Stevens, N. W. Ruktanonchai, A. J. Tatem, and M. C. Castro, "GridSample: an R package to generate household survey primary sampling units (PSUs) from gridded population data," *International Journal of Health Geographics*, 2017, *16*, 25.

Verwimp, P. and J. Mu noz Mora, "Returning Home after Civil War: Food Security and Nutrition among Burundian Households," *Journal of Development Studies*, 2018.

Voors, Maarten J., Eleonora E. M. Nillesen, Philip Verwimp, Erwin H. Bulte, Robert Lensink, and Daan P. Van Soest, "Violent Conflict and Behavior: A Field Experiment in Burundi," *American Economic Review*, April 2012, *102* (2), 941–64.

Wagenaar, B., O. Augusto, K. Asbjornsdottir, A. Akullian, N. Manaca, F. Chale, A. Muanido, A. Covele, C. Michel, S. Gimbel, T. Radford, B. Girardot, and K. Sherr, "Developing a representative community health survey sampling frame using open-source remote satellite imagery in Mozambique," *International Journal of Health Geographics*, 2018.

Wardrop, N. A., W. C. Jochem, T. J. Bird, H. R. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson, S. Juran, V. Seaman, and A. J. Tatem, "Spatially disaggregated population estimates in the absence of national population and housing census data," *Proceedings of the National Academy of Sciences*, 2018, *115* (14), 3529–3537.

Weber, E. M., V. Y. Seaman, R. N. Stewart, T. J. Bird, A. J. Tatem, J. J. McKee, B. L. Bhaduri, J. J. Moehl, and A. E. Reith, "Census-independent population mapping in northern Nigeria," *Remote Sensing of Environment*, 2018, *204*, 786–798.

Wiens, K., P. Mawien, J. Rumunu, D. Slater, F. Jones, S. Moheed, A. Caflisch, B. Bior, I. Jacob, R. Lako, A. Guyo, O. Olu, S. Maleghemi, A. Baguma, J. Hassen, S. Baya, L. Deng, J. Lessler, M. Demby, V. Sanchez, R. Mills, C. Fraser, R. Charles, J. Harris, A. Azman, and J. Wamala, "Seroprevalence of Severe Acute Respiratory Syndrome Coronavirus 2 IgG in Juba, South Sudan," *Emerging Infectious Diseases*, 2021.

Yeh, C., A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke, "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa," *Nature Communications*, 2020.

Appendix

Table A1: Comparing Pros and Cons Across Sampling Frame Methods

Method	Pros	Cons
Full Populations Listing	Contains detailed demographic information More easily informs policy decisions	Difficult to obtain in conflict environments Often out of date in fragile regions Many regions lack the labor force necessary Logistically complex Costly Time delay Potential to attract population influx due to confusion with registration for humanitarian assistance Increased time in the field leads to higher possibility of danger
Spin-the-Pen	Convenient Cheap	Tendency for bias Time delay
Manual Tagging from High Res Imagery	Can be collaborative, such as with OpenStreetMap Enumeration can be done safely	Time Consuming It can be costly to obtain such imagery
Automated Tagging from High Res Imagery	Saves time Inexpensive Enumeration can be done safely	Potential for bias due to mismatch between training imagery and use case Must be reviewed Harder to identify mobile and displaced populations because they are not static across imagery

Note: Author's synthesis based on Aguilera et al. (2019); Bauer (2014); Escamilla et al. (2014); Galway et al. (2012); Kamedjeu (2009); Lin and Kuwayama (2016); Wagenaar et al. (2018).

Table A2: Building Status by Community Type

Building State	N	%
Urban		
Yes, residential building (including informal or temporary shelter)	659	83.52
No, commercial, industrial or administrative building	19	2.41
No, construction site (nobody living inside)	33	4.18
No, destroyed building	16	2.03
No, there is no building	21	2.66
No, abandoned house	41	5.20
Total	*789*	*NA*
Rural		
Yes, residential building (including informal or temporary shelter)	398	98.76
No, commercial, industrial or administrative building	1	0.25
No, destroyed building	3	0.74
No, there is no building	1	0.25
Total	*403*	*NA*

textbfBuilding Status by Community Type. Reports the condition of buildings visited by each survey team, split by GHS-POP community type (Urban: $\geq$300 inhabitants/km^2; Rural: $<$300 inhabitants/km^2). Percentages are unweighted shares within each community type.

Table A3: Building Status by Identification Method

	Original and Kept Points		Found Points	
Building State	N	%	N	%
Yes, residential building (including informal or temporary shelter)	823	85.37	238	91.19
No, commercial, industrial or administrative building	20	2.07	3	1.15
No, construction site (nobody living inside)	41	4.25	5	1.92
No, destroyed building	17	1.76	6	2.30
No, there is no building	27	2.80	4	1.53
No, abandoned house	36	3.73	5	1.92
Total	*964*	*NA*	*261*	*NA*

textbfBuilding Status by Identification Method. Reports the condition of buildings visited by each survey team, split by identification method. Original and Kept points are ML-generated footprints retained after validation; Found points were added manually via satellite imagery review. Percentages are unweighted shares within each identification method.

Table A4: Weighted vs. Unweighted Estimates: Respondent Types by Identification Method

	Original and Kept Points (%)	Found Points (%)	Difference (pp)
Weighted			
Host	23.01	12.20	-10.81***
IDP	17.84	24.87	7.03**
Returnee	59.14	62.92	3.78
Unweighted			
Host	28.64	16.95	-11.69***
IDP	20.62	27.54	6.93**
Returnee	50.74	55.51	4.77

textbfWeighted vs. Unweighted Estimates. *p<0.10, **p<0.05, ***p<0.01. Replicates Table 4 using unweighted OLS (Unweighted panel) alongside the exposure-weighted estimates reported in the main text (Weighted panel). Weights are the number of eligible buildings within 300 m of the respondent (textttexposure.n). Difference column reports the OLS coefficient on an indicator for manual identification (Found = 1, Original/Kept = 0).

Table A5: Community-Level Summary Statistics

Community	Type	Area (km^2)	Alg.				
Points	Found	Total	% Found	Res. Acc. (%)			
Al Teneraa	Rural	15.9	64	82	146	56.2	100.0
Al-Ajelyah	Rural	12.1	408	350	758	46.2	98.6
Al-Shehabbi	Rural	23.0	743	302	1,045	28.9	100.0
BZBZ	Rural	54.3	7,591	1,145	8,736	13.1	100.0
Rabeea and Samma	Rural	33.8	2,204	701	2,905	24.1	96.1
Yaramjah	Rural	3.1	1,533	536	2,069	25.9	100.0
8 Shibat	Urban	17.4	13,868	1,272	15,140	8.4	89.9
Al Forat	Urban	9.0	1,117	238	1,355	17.6	NA
Al Tajneed Qtr	Urban	2.9	1,289	740	2,029	36.5	88.4
Al-Obaidy 1	Urban	1.2	640	171	811	21.1	75.5
Duguri	Urban	9.2	4,279	226	4,505	5.0	NA
Gaza	Urban	4.0	736	195	931	20.9	93.8
Hay Al Askari	Urban	1.9	4,415	394	4,809	8.2	76.6
Hay Al Teen	Urban	1.9	2,634	197	2,831	7.0	82.7
Hay Tal Baajah	Urban	5.8	3,363	1,203	4,063	29.6	75.3
Markez Baaj	Urban	1.3	1,051	400	1,451	27.6	71.7
Nahda Sharqya	Urban	4.4	1,283	542	1,825	29.7	86.2
Rajm Hadid	Urban	2.7	2,594	883	3,477	25.4	100.0
Yangija	Urban	6.0	2,577	83	2,660	3.1	81.4

Community-Level Summary Statistics. Alg. Points: number of Microsoft GlobalML building footprint centroids within each community boundary. Found: manually placed centroids added from current Planet Labs imagery. Total: validated algorithmic points plus Found. % Found: Found as a share of Total. Res. Acc.: share of visited points that were occupied residential buildings. Urban/Rural classification based on GHS-POP population density ($\geq$300 inhabitants/km^2 = Urban).

Table A6: Respondent Composition by Identification Method, Split by Bombardment Intensity

	Original/Kept (%)	Found (%)	Difference (pp)
Low bombardment			
Host	15.05	12.67	-2.37
Returnee	75.98	54.62	-21.37***
IDP	8.97	32.71	23.74***
Medium bombardment			
Host	16.43	9.06	-7.38*
Returnee	73.50	72.97	-0.53
IDP	10.07	17.97	7.9**
High bombardment			
Host	40.98	26.32	-14.66*
Returnee	20.40	41.83	21.42***
IDP	38.62	31.85	-6.77

Note: Replicates Table 4 within terciles of ACLED *aerial and artillery* bombardment intensity. Event density is computed as Air/drone strike and Shelling/artillery/missile attack events (ACLED structured sub_event_type) per km^2 within a 5km buffer of the community centroid, 2016–2023. IEDs, car bombs, suicide bombs, grenades and chemical-weapon events are excluded by construction. Communities ranked by aerial/artillery events per km^2 and split at tercile boundaries. Difference column reports the exposure-weighted LPM coefficient on the Found indicator, estimated separately for each respondent type within each tercile. *p<0.10, **p<0.05, ***p<0.01.

Table A7: Community-Level Bombardment, Shelter, and Detection Summary

Community	Urban/Rural	N	Found (%)	IDP (%)	Returnee (%)	Tent (%)	Informal (%)	Destroyed (%)	ACLED aerial/km^2	UCDP aerial/km^2	UCDP aerial (count)
Rajm Hadid	Urban	53	13.2	34.0	5.7	0.0	0.0	24.5	310.15	15.16	41
8 Shibat	Urban	89	6.7	52.2	21.7	0.0	2.2	52.8	16.01	1.15	20
Hay Al Askari	Urban	47	8.5	30.6	11.1	0.0	6.4	8.5	32.25	1.04	2
Hay Al Teen	Urban	52	7.7	37.2	20.9	0.0	19.2	28.8	31.90	1.03	2
Markez Baaj	Urban	92	28.3	23.1	53.8	0.0	0.0	8.7	34.08	0.76	1
Gaza	Urban	80	15.0	2.7	70.7	0.0	0.0	38.8	0.00	0.25	1
Al Tajneed Qtr	Urban	86	41.9	21.3	77.3	0.0	0.0	82.6	9.27	0.00	0
Al Teneraa	Rural	31	58.1	6.5	90.3	0.0	0.0	71.0	0.00	0.00	0
Al-Ajelyah	Rural	73	21.9	6.9	55.6	0.0	1.4	38.4	0.00	0.00	0
Al-Obaidy 1	Urban	53	18.9	12.5	67.5	0.0	5.7	18.9	6.51	0.00	0
Al-Shehabbi	Rural	72	27.8	4.2	69.4	0.0	0.0	52.8	0.35	0.00	0
BZBZ	Rural	65	53.8	83.1	0.0	78.5	78.5	66.2	0.00	0.00	0
Hay Tal Baajah	Urban	93	19.4	10.0	77.1	0.0	0.0	19.4	14.74	0.00	0
Nahda Sharqya	Urban	58	20.7	2.0	46.9	0.0	0.0	39.7	2.50	0.00	0
Rabeea and Samma	Rural	103	27.2	9.1	81.8	0.0	0.0	68.9	0.21	0.00	0
Yangija	Urban	86	4.7	2.9	81.4	0.0	0.0	60.5	0.17	0.00	0
Yaramjah	Rural	59	3.4	49.2	8.5	0.0	0.0	39.0	18.34	0.00	0

Note: Each row is one study community. Found (%) = share of visited buildings identified via manual satellite review rather than algorithmic footprint. Tent and Informal columns report the share of respondents living in tents or informal shelters (tent, makeshift, prefab/caravan). Destroyed (%) = share reporting house destroyed since 2014 (still destroyed or rebuilt). **Both bombardment densities cover aerial and artillery events only.** ACLED aerial/km^2 counts Air/drone strike and Shelling/artillery/missile attack events (ACLED structured sub_event_type) within a 5km buffer of the community centroid, 2016–2023, normalised by area. UCDP aerial/km^2 counts UCDP GED events classified as aerial/artillery via regex on source headlines and articles (with IED/car-bomb/suicide-attack veto), within the same buffer, 2014–2017, normalised by area. UCDP aerial (count) reports the same events in absolute terms. IED, car-bomb, suicide-bomb, grenade and chemical-weapon events are excluded from both measures by construction.