Leveraging geospatial techniques and publicly available datasets to develop a cost-effective and digitized national sampling frame in Armenia

Saida Ismailakhunova
Avralt-Od Purevjav
Tsenguunjav Byambasuren
Sarchil Qader

April 11, 2025

Abstract

The absence of a reliable and accurate national sampling frame represents a significant methodological constraint in conducting representative national surveys. This limitation undermines policy and research efforts in many developing countries, particularly those facing substantial internal displacement and relocation due to territorial challenges and conflicts. This paper addresses this challenge by developing Armenia's first digitized national sampling frame—a country where reliable and accessible sampling frames for household and individual surveys are severely constrained. The study begins by reviewing existing national sampling frames and highlighting their scientific and logistical limitations. It then proposes efficient tools, geospatial techniques, and datasets for developing urban and rural classification suitable for surveys, as well as digitized pre-census enumeration areas, which can be used as a national sampling frame. The proposed methods and strategies offer several innovations and advantages over traditional approaches. First, the process of creating a digitized national sampling frame is fully automated. As a result, the digitization of Armenia's pre-census enumeration areas was completed in less than three months with limited resources, whereas a manual process would have taken years and required significant financial investment. Second, all input datasets were publicly available, which is crucial for scaling the method to other countries. Third, because the process is computer-based, the resulting output is free from the geometric errors often associated with manual methods. Fourth, the population parameter is derived from gridded population data, which accounts for recent urban changes and migration, thereby maximizing the representativeness of the results. The results show that the urban-rural classification population total strongly correlates with the 2011 census outputs. The pre-enumeration area (preEA) boundaries align with international standards, including nesting within administrative boundaries and aligning with visible ground features such as roads, rivers, and infrastructure. This new sampling frame was successfully applied to the World Bank's "Listening to Armenia" survey, showcasing its potential for other socioeconomic surveys in the country. Furthermore, the method can be utilized to efficiently generate and update national sampling frames in other countries.

Keywords: Sampling frame, Enumeration areas, Census, Geospatial, Household surveys, Developing countries

1 Introduction

A national sampling frame, which includes primary sampling units (PSUs) such as enumeration areas derived from a recent census, is vital for collecting reliable and representative data. However, national sampling frames face several critical challenges globally, especially in developing countries and conflict-affected regions where conducting representative surveys is crucial for high-quality research and policy analysis with minimal bias. Many countries, including the Republic of Armenia, rely on national sampling frames based on census enumeration areas from previous population censuses. These frames, however, are often outdated, non-digital, incomplete, and difficult to access, limiting their use to local government organizations while excluding researchers, academics, and international organizations. This issue is particularly acute in Armenia, where the recent large-scale refugee crisis and significant shifts in population distribution due to the Nagorno-Karabakh conflict have further compounded these challenges (Gale et al. 2023). Consequently, the lack of accessible, up-to-date population sampling frames undermines efforts to conduct surveys for statistical, policy, and research purposes.

The digital national sampling frame is typically based on three sources: census enumeration areas (EAs), sub-national administrative boundaries, and a gridded population sampling frame. A census is conducted every 5 to 10 years, depending on financial and administrative costs, and the nature of the census questions. For instance, Armenia's sampling frame relies on the 2011 Census, which is now severely outdated due to demographic changes resulting from population displacements and mobility across regions caused by territorial conflicts. In 2022, the Committee of the Republic of Armenia (ArmStat) conducted a Population Census using a combined approach of administrative data and a sampled census, offering a potential up-to-date national sampling frame (ARMSTAT, 2022). However, due to the COVID-19 pandemic, the census was conducted in a hybrid format, as was the case in many countries. This frame, based on a sample of 25% of the addresses in the State Population Register (SPR), is restrictive and largely inaccessible (ARMSTAT, 2022). Several attempts have been made to access the State Population Register (SPR) but failed to obtain the sampling frame based on SPR addresses. Since this dataset is not readily available, it is essential to find an alternative methodology to update the data from the previous census.

One of the early techniques to address this challenge is grid sampling, where cells—rectangular areas with population estimates—serve as PSUs. The sampling frame relies on millions of grid cells, which are publicly available from various data sources, such as WorldPop (WorldPop, 2025), Geo-Referenced Infrastructure and Demographic Data for Development (GRID3) (GRID3, 2025), Global Human Settlement Layer (GHS-POP) (Global Human Settlement Layer, 2025), Gridded Population of the World version 4 (GPWv4) (International Union for the Scientific Study of Population Toggle menu, 2025), High Resolution Settlement Layer (HRSL) (Meta and CIESIN, 2025), LandScan Global (Lebakula et al., 2024), LandScan HD (Weber et al., 2017), accessible through platforms like Google Earth Engine. The grid size varies depending on the data source, and the selection of grid size and data provider can differ from one country to another. However, grid sampling comes with several challenges. First, grid boundaries are often unnatural, cutting through buildings and disregarding visible geographic features (Qader et al., 2021). Second, although the spatial size of grids is uniform, the population size within each grid can vary significantly. As a result, sparsely populated grids may need to be collapsed, while densely populated grids require segmentation (Qader et al., 2020, 2021). Over recent decades, several methodological approaches and tools have been developed to create gridded population sampling frameworks (Cajka et al., 2018; Qader et al., 2020; Thomson et al., 2020).

Other researchers and surveyors have utilized other geospatial techniques and datasets to develop various national sampling frames tailored to their specific needs and objectives. Kassié et al. (2017) outline a sampling protocol for a health survey in Bobo-Dioulasso, Burkina Faso, using urban typology based on infrastructure and satellite imagery. The method surveyed 1,045 households, providing an alternative approach for areas with limited data. In the context of a hard-to-reach and mobile population, a random geographic cluster sample (RGCS) was explored to address undercoverage in household surveys in Ethiopia, by selecting random points and interviewing all eligible respondents within designated circles (Himelein et al. 2014). A community-based survey was conducted using area-based stratified random sampling and geospatial technology to examine social determinants of health and their association with obesity prevalence among Hispanics and non-Hispanic whites in a rural Southeastern U.S. community (Howell et al. 2020). The lack of translation of these methods into user-friendly tools, along with challenges in their reproducibility in certain contexts, presents difficulties in replicating these methods in other countries, especially in regions where geospatial skills are limited. Therefore, enhancing geospatial capacity and developing user-friendly tools is crucial to fully leverage geospatial techniques, ensuring the creation of more accurate and representative sampling frames.

This paper begins by systematically assessing the existing sampling frames in Armenia, with a particular focus on those utilizing census settlements—villages in rural areas, towns in urban areas, districts in the capital city, Yerevan—and electoral precincts as primary sampling units (PSUs). It evaluates the strengths and limitations of these sampling frames, offering potential solutions to address the challenges and shortcomings identified. To overcome the limitations of the existing national sampling frames, this paper proposes an innovative national sampling frame based on pre-census enumeration areas (pre-EAs), employing a novel approach alongside the most recent population estimates.

This paper presents Armenia's first digital national sampling frame, successfully developed using a range of innovative geospatial techniques and datasets. Additionally, multiple datasets and maps have been produced to support ground survey data collection. The proposed national sampling frame, where primary sampling units (PSUs) or enumeration areas (EAs) are automatically created, offers several advantages over traditional sampling frames (Qader et al., 2021). Firstly, the EAs are designed with natural boundaries such as streets and rivers, ensuring they align with geographic features. Secondly, administrative boundaries are strictly adhered to, as the EAs are nested within these boundaries by design. While some automatically generated boundaries can be unnatural, they are not cutting houses and structures, and the automatic creation of EAs minimizes geometric errors (such as pockets, disjoint sections, and overlaps), making the process significantly more resource-efficient compared to manual creation by cartographers. Thirdly, the population estimates within the EAs are generally homogeneous, avoiding extremes in size, although some further adjustments may be necessary. Although basic comparisons were made to assess gridded population outputs, a comprehensive validation of gridded population estimates is outside the scope of this paper due to limitations in data, time, and resources.

The datasets used to create the semi-automatic mapping of pre-EAs include high-resolution gridded population data, the spatial distribution of settled areas, and publicly available natural and administrative boundaries from various sources, such as OpenStreetMap (OSM) and WorldPop. Cross-validation exercises have been carried out to assess the applicability of the proposed sampling frame and compare it with other existing frames in the country. These datasets include: (i) the 2011 Census with spatial information on regions (marzes) and settlement types (urban or rural), (ii) census settlements based on the 2011 Census, (iii) 2023 electoral precincts, and (iv) additional population data from ArmStat at the aggregate level, including marz-settlement type (urban and rural regions).

This paper makes several contributions to the literature and the field of survey sampling. First, it presents a national sampling frame for Armenia based on pre-census enumeration areas (pre-EAs), demonstrating the applicability of a semi-automated spatial technique that could benefit other countries. The method has been implemented and tested in countries such as Somalia, which lacks a digital national sampling frame (Qader et al., 2020, 2021), Cameroon, which requires a customized national sampling frame for refugees (Darin et al. 2024), the Democratic Republic of Congo (Qader et al., 2023), and Burkina Faso (Qader et al., 2022). However, this is the first application of the tool in Central Asia. Second, the national sampling frame developed in this paper contributes to survey data collection in Armenia. The use of multiple sampling frames in the country often makes it difficult for researchers, policymakers, and others to compare results across surveys. By providing a unified, standardized sampling frame based on publicly available datasets, this paper helps ensure consistency across surveys and avoids discrepancies in population estimates, offering a methodological contribution to the field. Third, the rigorous evaluation of various sampling frames, including the new national sampling frame, contributes to the survey sampling literature and practice in Armenia. To the best of our knowledge, no study has yet systematically compared various sampling frames in Armenia. This new sampling frame does not replace existing frames but can complement them, particularly the national census frame, offering an alternative approach. Finally, this work highlights the value of public datasets such as OpenStreetMap. The availability of high-quality, public geospatial data can generate substantial societal value, potentially amounting to tens or even hundreds of millions of dollars, even before considering indirect benefits (Hansen and Schrøder, 2019).

2 Existing and Accessible Sampling Frames

The household survey is one of the most widely used methods for gathering population and socioeconomic data (United Nations (UN) 2005); United Nations Children's Fund (UNICEF) 2012). To ensure the accuracy of these surveys, a reliable sampling frame is essential at the national level. These frames are typically established and updated during national housing and population censuses. However, some countries lack a digital national sampling frame due to non-digitized censuses or, in more extreme cases, may not even have a physical map of the census frame, or it may be inaccessible. Furthermore, these frames often become outdated because demographic factors evolve quickly, and most censuses are conducted every ten years.

In Armenia, there is no functional or accessible map or cartographic information that can be used for a national sampling frame, posing a significant barrier to conducting nationally representative socioeconomic surveys. In addition, the country does not currently have an up-to-date and digitized national sampling frame. The last census conducted in Armenia was in 2011 (ARMSTAT, 2024), and there are no up-to-date enumeration areas available for use as national sampling frames for representative socioeconomic surveys. The spatial resolution of the census data in use today is limited to the provincial or district level (2nd and 3rd administrative units), making it difficult to determine how people are distributed at finer scales, such as the facility, sub-district, or neighbourhood levels—where most policy interventions typically occur.

In developing countries, creating a sampling frame for surveys that includes representative community samples usually involves manually delineating small geographic areas (or enumeration areas) on high-resolution satellite imagery. While this method is commonly employed by National Statistical Offices (NSOs), it is logistically complex and requires substantial resources, including Geographic Information System (GIS) experts and extensive training (Qader et al., 2020, 2021). Additionally, this process is both time-consuming and expensive, often resulting in delays to the survey. For instance, it can take two to three years to complete a survey (Grosh and Glewwe, 1995). These challenges highlight the need for a faster and more cost-effective approach to sampling frame and population enumeration methodologies.

Before discussing the new sampling frame, an overview of the existing sampling frames in the country is provided. Three datasets have been identified as potential sources for developing a national sampling frame for household surveys: the 2022 Census with addresses from the State Population Register (SPR), the 2011

Census with settlement data (villages in rural areas, towns in non-Yerevan urban regions, and districts in Yerevan), and 2023 election data with electoral precincts.

The ArmStat conducted a Population Census in November 2022, employing a combined approach based on administrative data from the State Population Register (SPR) and a 25% sample of SPR addresses (ARMSTAT, 2022). In this dataset, primary sampling units (PSUs) correspond to the workload of SPR addresses assigned to each enumerator during the census. While these PSUs lack the identifiable boundaries of traditional census enumeration areas (EAs), they can still serve as a sampling frame since they cover approximately 93% of all addresses in the country. The household listings in this dataset were last updated in October 2022, although the addresses are distant due to the large size of the EAs. Despite its strengths, this frame is inaccessible, as the data is only available on a restricted computer at ArmStat (Pettersson, 2023).

Given these challenges, this paper focuses on the latter two datasets: the 2011 settlement-based frame and the 2023 electoral precinct-based frame. Their respective advantages and limitations are discussed in detail.

2.1 Census Settlements

The most granular spatial information available in this dataset is at the "settlement" level. There are 1,037 settlements (980 villages, 45 towns, and 12 districts in Yerevan). One of the key advantages of this sampling frame is that it identifies over 1,000 distinct geographical areas, which is more granular than simply using regions or marzes.

A key challenge with this dataset is its heterogeneous spatial coverage units. Suppose that 400 settlements were selected in the first stage of the two-stage stratified cluster sampling design as primary sampling units (PSUs). Given the large populations in the 12 districts of Yerevan, it is likely that all districts will be selected using a probability proportional to size (PPS) approach. If 10 households were randomly selected from each district in the second stage, the sample size from Yerevan would total 120 households, which represents only 3% of the total sample of 4,000 households. However, according to the 2011 Armenia Census, Yerevan accounts for approximately 35% of the population and 38% of the total households. One can select a disproportionate number of households from each PSU in the second stage to account for the variations in the size of the PSUs in the first stage. For example, selecting 100 households from each district in Yerevan would yield 1,200 households from Yerevan, which is 30% of the sample.

However, a notable drawback of this frame is the presence of large settlements, particularly in Yerevan. While identification information is unavailable, the Census frame from the Committee of the Republic of Armenia (ArmStat) includes approximately 12,000 enumeration areas, which means settlements in this frame are, on average, 12 times larger than the Census enumeration areas. Using large primary sampling units (PSUs) in this manner could undermine the integrity of the two-stage sampling design, effectively reducing it to a one-stage design. Large settlements must be subdivided into smaller and more manageable PSUs to resolve this. In the past, large PSUs have been manually segmented into smaller units, as demonstrated in Nepal (Central Bureau of Statistics of Nepal, 1996); however, this traditional approach is both costly and time-consuming. This paper proposes an innovative technique for dividing these large areas into smaller, more practical units.

Another issue with this potential sampling frame is that the population data based on the 2011 Census is outdated and misallocated. While outdated data typically is not a major concern for national sampling frames (since any survey conducted before the 2022 Armenia Census could use the 2011 Census frame), it presents a more significant problem in Armenia, where household displacement and domestic migration due to territorial conflicts have been substantial in recent years (Gale et al. 2023). As a result, the current population distribution may differ significantly from that recorded in the 2011 Census. To address this, population data can be updated using population growth rates at more aggregate levels. If updates are made at a finer level than administrative units, the population distribution across areas can be adjusted to better reflect the current reality. However, any adjustments at the administrative unit level or more aggregate levels, such as regions, would not alter the probability of a PSU being selected in the first stage of the PPS process. Since PPS selection is based on administrative units, any monotonic transformation of PSU size within these units would not affect the selection probability. Therefore, adjustments to population and household data should be made at a level more granular than administrative units to ensure the integrity of the sampling process.

2.2 Electoral Precincts

The confidential microdata on electoral precincts is originally sourced from the Central Election Committee of Armenia. There are two main advantages to using this dataset as a sampling frame for household surveys. First, the data is regularly updated and reflects current information. Second, with 1,992 electoral precincts, the dataset exceeds the 1,037 settlements in the 2011 Census settlement data. As a result, the spatial information is more granular than that provided by the 2011 Census, and issues related to a few large-sized PSUs are less pronounced compared to a sampling frame based on census settlements. However, similar to the "settlements" in the 2011 Census data, using electoral precincts as primary sampling units (PSUs) also presents challenges related to large-sized PSUs. It is important to note that there are also smaller electoral precincts, which are less problematic. As noted in Pettersson (2023), the size of voting point areas in Armenia ranges from 7 addresses to 1,200 addresses. Smaller voting point (VP) areas can be merged with neighbouring areas, while larger VP areas can be divided into smaller segments. The process of merging smaller VP areas should be relatively straightforward, but segmenting large areas may incur additional costs, as it requires spatial analysis and likely some cartographic work.

Additionally, the boundaries of primary sampling units (PSUs) are crucial to ensure that enumerators do not exceed the targeted area. This feature is lacking in both sampling frames discussed in this section, as no boundaries (neither digital nor physical) are available for the Census settlements or electoral precincts. However, this is a more significant issue for the precinct-based sampling frame, as precincts are relatively smaller in size compared to settlements. As a result, the likelihood of enumerators inadvertently entering neighbouring, non-selected PSUs is higher for precincts. In the case of very small electoral precincts, enumerators may stray outside the designated area if they are not provided with proper maps during fieldwork.

The advantages and disadvantages of a sampling frame based on the 2023 electoral precincts indicate that it is relatively more favourable than the frame based on the 2011 Census settlements. Consequently, the 2023 electoral precincts have been further evaluated as a sampling frame for representative individual- and household-level surveys, with an exploration of the data on electoral precincts. The size of each electoral precinct is measured by the number of voters or the adult population, excluding children or individuals under 18 years old. Table 1 illustrates the distribution of strata size using both the total and adult populations. The stratum is defined as a combination of marz and settlement type—urban or rural status, as seen in other official surveys like Armenia's Demographic and Health Survey (DHS) (National Statistical Service et al., 2017). The 2022 population data at the strata level is sourced from the Committee of the Republic of Armenia

(ArmStat).[1] The total population in 2022, as shown in Column 1, is 2.977 million. Column 2 displays the 2023 number of voters (adult population aged 18 or older) based on the electoral precinct-based sampling frame. Column 3 shows the difference between the total population and the adult population over subsequent years. Although the two population figures correspond to different years, some irregularities are observed, such as the adult population exceeding the total population by approximately 11,000 people in rural areas of Lori.

Table 1: Population and number of voters across strata

Marz name	Settlement type	Strata ID	Population (2022 ArmStat data)	Number of voters (18 or older, 2023 election data)	Difference between total and adult population
			(1)	(2)	(3) = (1) - (2)
Aragatsotn	Urban	1	26,738	27,847	-1,109
Aragatsotn	Rural	2	98,949	82,622	16,327
Ararat	Urban	3	72,294	56,730	15,564
Ararat	Rural	4	186,983	150,805	36,178
Armavir	Urban	5	82,953	76,205	6,748
Armavir	Rural	6	183,703	137,823	45,880
Gegharkunik	Urban	7	65,902	61,823	4,079
Gegharkunik	Rural	8	162,809	109,803	53,006
Kotayk	Urban	9	137,493	126,680	10,813
Kotayk	Rural	10	116,364	94,876	21,488
Lori	Urban	11	124,050	134,962	-10,912
Lori	Rural	12	87,532	76,157	11,375
Shirak	Urban	13	133,620	128,691	4,929
Shirak	Rural	14	96,856	77,832	19,024
Syunik	Urban	15	90,205	65,044	25,161
Syunik	Rural	16	44,350	33,042	11,308
Tavush	Urban	17	49,859	39,214	10,645
Tavush	Rural	18	69,943	58,907	11,036
Vayots Dzor	Urban	19	16,160	16,811	-651
Vayots Dzor	Rural	20	31,501	25,678	5,823
Yerevan	Urban	21	1,098,866	824,317	274,549
Armenia			2,977,130	2,405,869	571,261

Notes: Column (2) presents the precinct data aggregated at the strata level. The 2023 election data on the number of voters or adult population and other spatial information are sourced from the Central Electoral Commission of Armenia.

As shown in Column 2 of Table 1, the total number of voters or the adult population is 2.405 million, which is quite close to the total population. This suggests that approximately 19% of the population is composed of children under 18 years old. However, other datasets indicate that children under 18 make up around 23-24% of Armenia's population. To further investigate this, the total adult population across various official data sources was examined for comparison. Table 2 presents the findings. According to the 2011 Armenia Census, the adult population share (aged 18 and older) is 77%, while the adult population share derived from a combination of the 2022 World Bank data (for the 0-14 age group) and the 2011 Armenia Census (for the 15-17 age group) is 76%. This suggests that the election data overestimated the adult population by approximately 4-5%. Despite these discrepancies, the PSU size based on the number of adults or voters does

[1] See https://armstat.am/file/doc/99538403.xlsx for the urban population and https://armstat.am/file/doc/99538413.xlsx for the rural population data at the region level in Armenia over time.

not pose a significant issue, as the total and adult populations across strata or administrative units are strongly and positively correlated, with a correlation coefficient of ϱ = 0.996 (p-value = 0.000).

Table 2: Share of adult population in different datasets

	Adult population (18 or older, 2011 Armenia Census)	Adult population (18 or older, 2022 World Bank data and 2011 Armenia census)	Number of voters (18 or older, 2023 election data)
	(1)	(2)	(3)
Share in total population	77%	76%	81%

Notes: The 2011 Census data used in Column 1 reports age-specific population, and the share of the population with 18 or older is shown.

In addition to the absolute value of PSU size, the distribution of the size measure across PSUs is also crucial. Figure 1 illustrates the distribution of the 2023 adult population across electoral precincts. Ideally, PSUs should be of equal size, or the PSU sizes should be evenly distributed across the sample frame. Population data from census frames typically follows a normal distribution, with few very small or large PSUs. However, the distribution of voters across electoral precincts in this case is U-shaped. The precinct size ranges from 10 to 2,061 voters, with a mean size of 1,208 and a median size of 1,399. This distribution highlights the need for merging and segmentation to make the electoral precinct-based frame more workable, aligning with the distribution of households observed in Pettersson (2023).

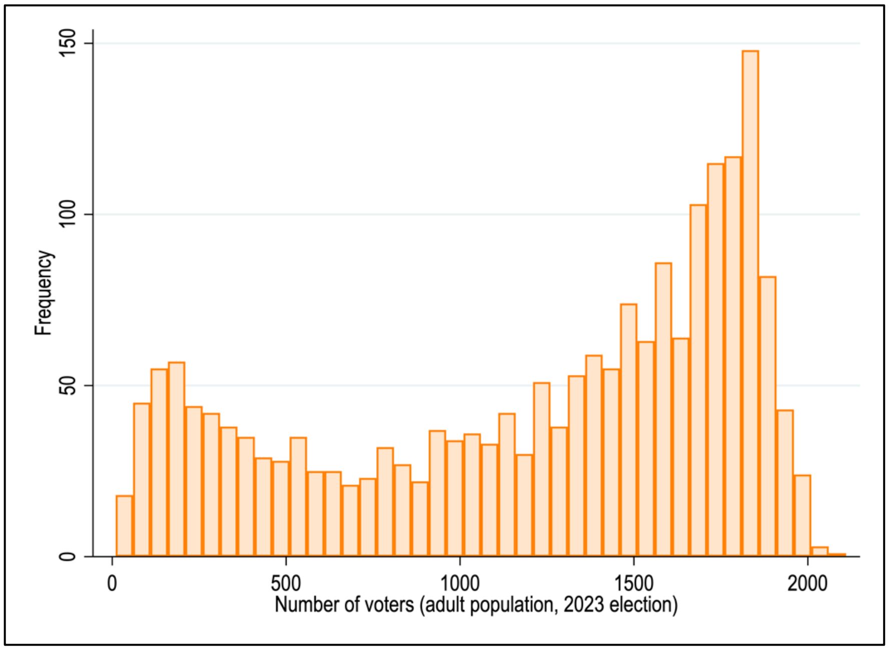

Figure 1: The distribution of voters or adult population across electoral precincts in Armenia in 2023

Finally, major and official surveys, such as the Demographic and Health Survey (DHS) for Armenia, rely on sample frames based on enumeration areas, rather than electoral precincts. Table 3 provides a summary of the sampling frames used in major surveys across Armenia. The absence of a usable and accessible sample frame based on enumeration areas highlights the need for a workable and up-to-date national sampling frame in the country. Therefore, given the limitations of the existing sample frames, this paper proposes the development of a national sampling frame based on pre-census enumeration areas (pre-EAs) in Armenia, which offers a conceptual improvement over the existing sampling frames evaluated in this section.

Table 3: Sampling frames of major surveys in Armenia

Surveys	PSUs in the sampling frames
Demographic and Health Survey (DHS)	Enumeration areas in the Armenia Population and Housing Census
Integrated Living Conditions Survey (ILCS)	Population census enumeration areas
UNICEF Multiple Indicator Cluster Survey (MICS, first ever in Armenia)	Currently in search of a suitable sampling frame

Notes: The PSUs stand for primary sampling units.

3 Data and Methods

The preEA production process involves several stages. First, various geospatial datasets were obtained and pre-processed. Next, urban and rural areas across the country were classified. The preEA tool was then applied to generate the preEA boundaries. Finally, both automatic and manual processes were used to post-process the preEA boundaries and validate them. Figure 2 illustrates the overall process involved in producing the national sampling frame in Armenia for this study.

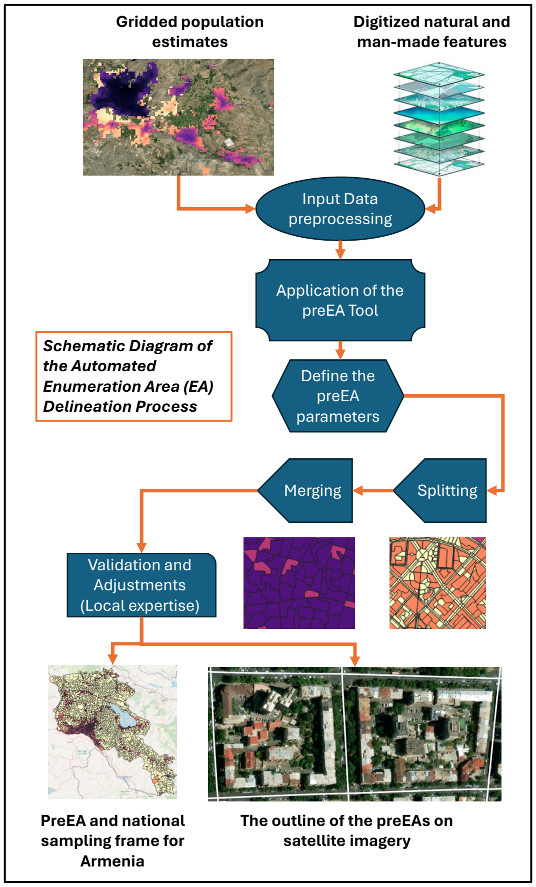
Gridded population estimates
Digitized natural and man-made features
Input Data preprocessing
Application of the preEA Tool
Schematic Diagram of the Automated Enumeration Area (EA) Delineation Process
Define the preEA parameters
Merging
Splitting
Validation and Adjustments (Local expertise)
PreEA and national sampling frame for Armenia
The outline of the preEAs on satellite imagery

Figure 2: Schematic diagram of the semi-automated enumeration area (EA) delineation process in Armenia

3.1 Input Datasets

This section outlines the datasets sourced from various organizations to establish the national sampling frame and facilitate field data collection in Armenia.

3.1.1 Gridded Population

The gridded population data for Armenia was obtained from WorldPop (Bondarenko et al., 2020) and is based on the 2020 population census or projection-based estimates for that year. This dataset provides an estimated total population per grid cell (shown in panel (a) of Figure 3). The data is available at a resolution of 3 arcseconds (approximately 100 meters at the equator) and can be downloaded in GeoTIFF format with the Geographic Coordinate System, WGS84 projection. The population estimate is represented in the units of one pixel. Regions marked with "NoData" indicate areas classified as unpopulated according to the Built-Settlement Growth Model (BSGM) developed by Nieves et al. (2020). The WorldPop gridded population dataset was generated by disaggregating projected subnational population totals into grid cells using machine learning techniques, incorporating various geospatial layers derived from satellite imagery (Stevens et al., 2015).

3.1.2 Digitized Features Visible from the Ground

The boundaries of enumeration areas (EAs) should align with prominent visible ground features to facilitate effective ground-based data collection. To ensure that the pre-EA boundaries meet these criteria during the automatic creation of the national sample frame in Armenia, extensive digitized ground features are required. These features, both natural and human-made, are primarily sourced from OpenStreetMap (OSM) (Geofabrik 2025). Panel (b) of Figure 2 displays information from the OSM dataset, including road networks, waterways, and railways. This data offers modifiable and updatable inputs, allowing for multiple iterations of EA generation. The figure illustrates the input datasets, where the two datasets have not yet been combined. Subsequent figures demonstrate how these input datasets are used to divide the country, along with the estimated total population for each area.

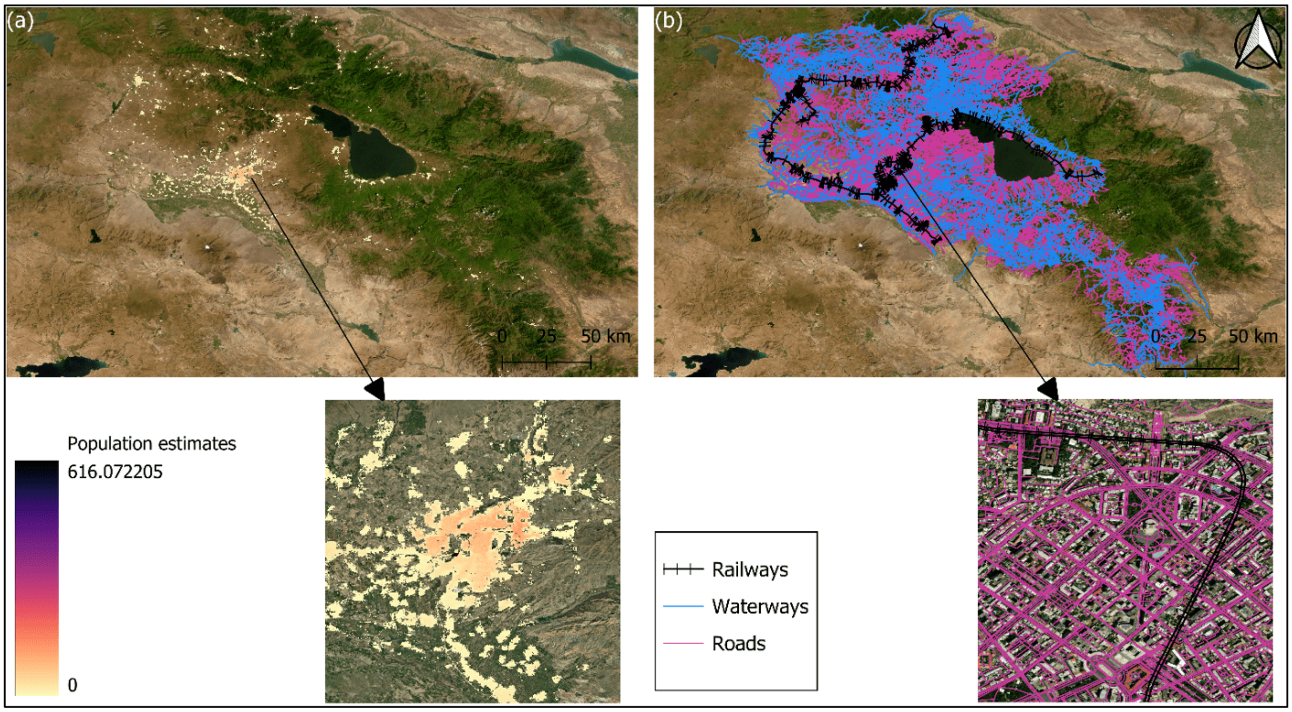

Figure 3: Population estimates and digitized visible ground features

Notes: Panel (a) shows the gridded population estimates at ~100×100 meters, while panel (b) depicts the digitized visible ground features from OSM. Basemap: ESRI Satellite Imagery.

3.1.3 Settlement Boundary and Classes

Determining the precise location and boundaries of each settlement is crucial for creating pre-EAs and guiding field operations. Accurate settlement boundaries help define settled areas and prevent the mixing of pre-EAs among large, populated areas. To achieve this, settlement boundaries were created, and settlements were classified into urban and rural categories. First, using the Global Human Settlement Layers (GHSL) product, settlements across the country were identified and converted into a vector layer to define the settlement boundaries. Subsequently, the settlements were classified into urban and rural categories with the assistance of GHSL classification. The GHSL provides a variety of settlement-related data in different spatial and temporal resolutions, along with multiple informative classes. The following two datasets from GHSL were used for settlement delineation and urban and rural classification.

GHS-BUILT-S R2023A: The spatial raster dataset, GHS-BUILT-S, illustrates the distribution of built-up (BU) surface estimates in five-year intervals from 1975 to 2030, along with two functional use components: the total BU surface and the non-residential (NRES) BU surface (Schiavina et al., 2023). Panel (a) of Figure 4 displays this data, which is generated by spatially and temporally interpolating five observed sets of multi-sensor and multi-platform satellite images, including those from Landsat and Sentinel 2.

GHS-SMOD R2023A: Settlements have been globally delineated and classified using the GHS Settlement Model layers (GHS-SMOD), which apply a logic based on cell cluster population size, population density, and built-up area densities, as recommended by the United Nations Statistical Commission and defined in Stage I of the Degree of Urbanization (EUROSTAT, 2021). The built-up surface, land layer, and a 1 km² population spatial raster dataset serve as inputs for the GHSL SMOD. As shown in panel (b) of Figure 4, the GHS SMOD classifies the 1 km² grid cells into three spatial entities at the first hierarchical level: "urban centre," "urban cluster," and "rural grid cells" (Schiavina et al., 2023).

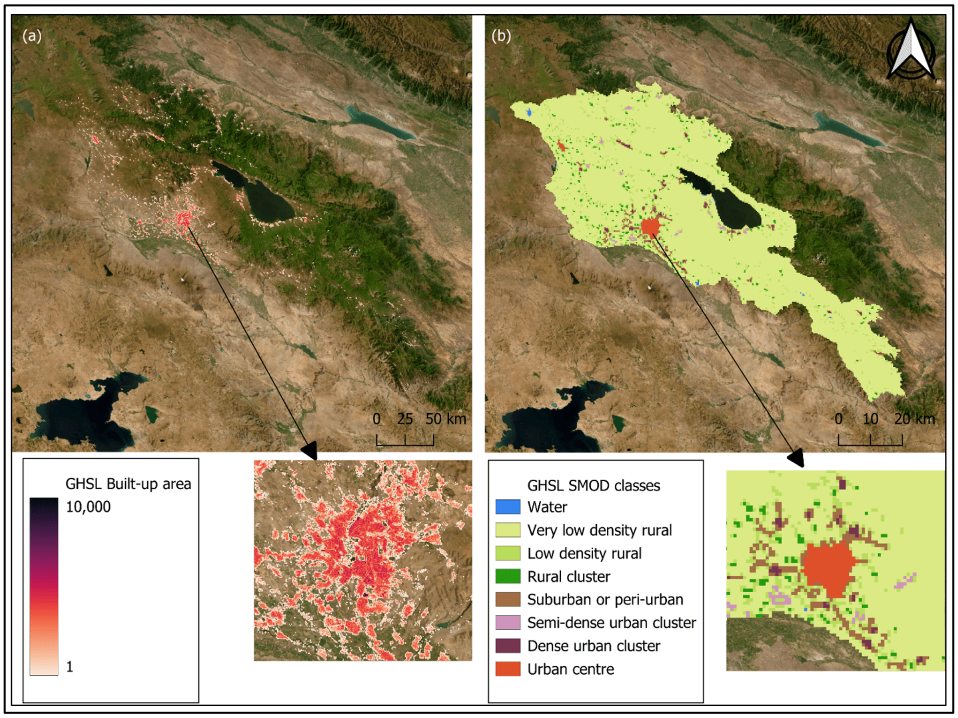

Figure 4: Global human settlement layer

Notes: Panel (a) presents the built-up area, and panel (b) shows the SMOD classes. Basemap: ESRI Satellite Imagery.

3.1.4 Administrative Boundary

It is essential to nest the produced enumeration area (EA) boundaries within administrative borders. The generated EA boundaries should be contained within these administrative boundaries. The necessary administrative border data for Armenia, based on the 2011 census, was sourced from the HDX website (HDX 2025).

3.2 Semi-automatic Creation of pre-EAs

This section outlines the semi-automated process used to generate the first digital national sampling frame in Armenia. While the national sampling frame is primarily created through automation, some manual adjustments may be necessary to enhance the outputs due to issues with inadequate or poor-quality input datasets. The process is divided into three main parts, which are detailed in the following sections.

3.2.1 Urban and Rural Classification

People living in and around cities are referred to as the urban population, typically characterized by a high population density. In contrast, the rural population is spread over large areas of land, predominantly found in developing regions. The overall population and geographic area are key factors in determining the size of enumeration areas (EAs). Achieving a balance between population and area constraints is essential to create EAs of manageable size. Due to the significant differences in population density and distribution between urban and rural administrative units, distinct criteria should be applied when forming EAs. Therefore, it is crucial to define urban and rural boundaries before creating pre-EAs, especially if the data is not readily available.

Armenia does not have well-defined boundaries to separate urban and rural areas. The following approaches have been used to establish the border between Armenia's urban and rural classes as a first step in developing the national sampling frame:

1. Take the GHS-SMOD dataset and extract all urban classes; combine them into a single class.
2. Convert the raster dataset from the combined classes to polygon vector format. These polygons represent the urban area.
3. Extract the built-up area with values greater than 0 from the GHS-BUILT-S data.
4. Create polygons from the raster built-up area. The country's settlement boundaries and extent are represented in this output.
5. Apply a 50-meter buffer to output 4 to account for recent urban growth and prevent cutting structures at the edge of the settlements.
6. Intersect output 5 with the administrative boundary in Armenia.
7. Make the necessary manual and automatic adjustments when necessary. For example, an administrative boundary should be regarded as urban if more than 90% of areas are urban (e.g., Yerevan).
8. Any polygon in output 7 that crosses the output 2 boundary is an urban area. The rural class encompasses the remaining portion of the administrative boundaries.

3.2.2 Application of the preEA tool

The “preEA” is a powerful and flexible tool developed by WorldPop in close collaboration with GeoData, with input from various governments, UN agencies, and global experts (Qader et al., 2021, 2023). As its development is ongoing, the tool is not yet publicly available. It is designed as a user-friendly QGIS plugin, built using the Python programming language. The implementation of the preEA tool is fully automated; however, certain preparatory processing steps are required before using the tool. These steps will be explained in the following sections.

Input Data Preparation: The data preparation consists of three steps. First, the input boundary datasets are reprojected into the projected UTM WGS 1984 coordinate system. The preEA tool is compatible with such projection to ensure that the output units are in familiar area units such as meters or feet. Second, the digitized boundary (roads, waterways, and railways) is masked by the extent of the administrative boundary. Third, to prevent the creation of sliver polygons in the outputs, double lines (such as motorways) are merged in the road datasets. It was accomplished by utilizing a 25-meter distance on Merge Divided Road in ArcGIS Pro. A single line will be created from any roads that fall within 25 meters of each other and have the same road code or class (Figure 5). The entire process was automated using ArcGIS ModelBuilder.

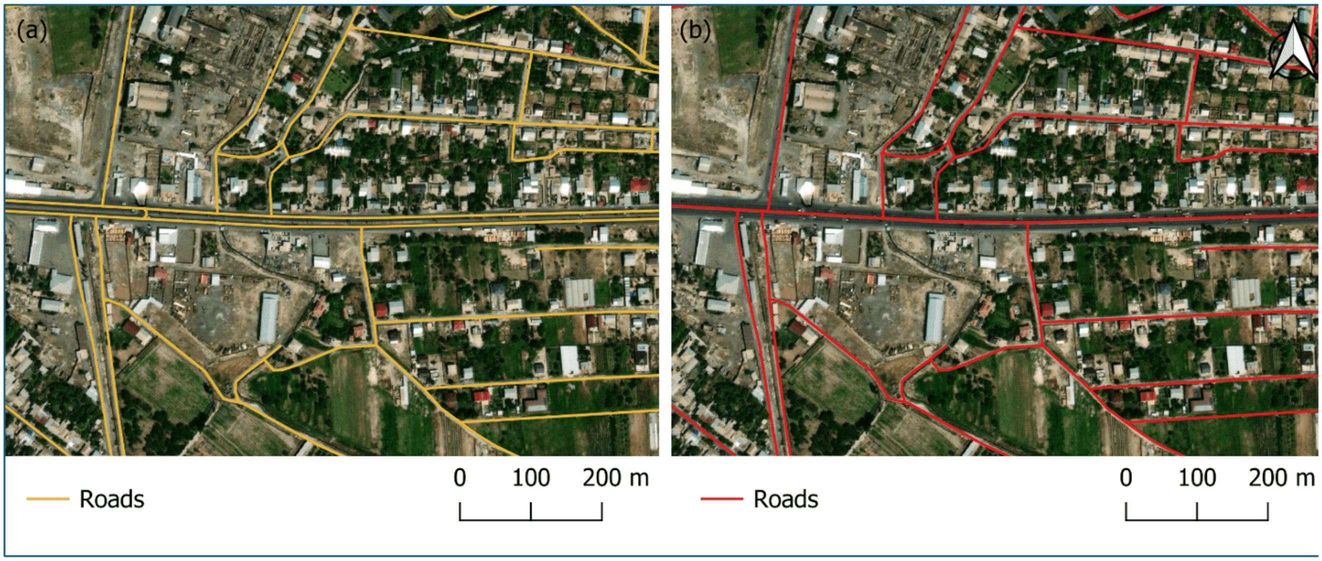

Figure 5: Preparation of road dataset

Notes: Panel (a) shows the original road data, while panel (b) presents the road data after applying the Merge divided road technique. Basemap: ESRI Satellite Imagery.

Create Uncrossable Features: Enumerators should avoid crossing major obstacles during data collection to enhance efficiency and reduce costs. To achieve this, certain features were designated as uncrossable:

(1) Uncrossable lines: In the OSM line features, the class of the digitized features is recorded in the “fclass” column. After a thorough visual assessment, road classes such as primary, secondary, trunk, and tertiary, as well as waterway classes like rivers, were extracted, combined, and designated as uncrossable lines.

(2) Uncrossable settlement boundary: It is preferable to keep enumeration areas (EAs) from different cities, towns, or villages separate to prevent the mixing of household and administrative hierarchies. To achieve this, three procedures were employed to establish uncrossable settlement boundaries. First, the Zonal Statistical technique was used to summarize the total population within each settlement polygon created in section 3.2.1. Second, a visual inspection approach was applied to identify the minimum population threshold for defining the uncrossable settlement boundaries. As a result, all settlement boundaries with a population exceeding 200 were designated as uncrossable. Third, the polygons containing more than 200 people were converted into lines and merged with additional uncrossable lines to generate the final set of uncrossable features.

Implementing the preEA Tool: In the first part of the process, the preEA tool divides the region into small geographic units by polygonising all the input feature datasets. The user must define hard constraints, including the maximum and minimum population size and geographic area. In addition to these hard constraints, it is essential to input administrative boundaries and uncrossable features that the pre-EA boundaries must adhere to. The user should also define soft constraints, such as the minimum length of shared boundaries and various weighting coefficients, to ensure that the generated pre-EAs meet both user expectations and global standards (Table 4).

Table 4: Parameter calibration in the preEA tool

Class	Maximum population	Maximum area	Minimum length shared boundaries	Area (coefficient)	Population (coefficient)	Share factor (coefficient)
Urban	1000	10km^2	20%	0	1	2
Rural	800	10km^2	20%	0	1	2

Once all the parameters are established, the small geographic units are merged until one of the hard constraints is satisfied. The output is in vector format, and each generated pre-EA includes necessary attribute information, such as the administrative boundary, total estimated population, area, and unique IDs (Random IDs and ID numbers generated using the serpentine technique). Table 4 presents the parameter calibration used in the preEA tool. The user can adjust priority parameters during the merging process. For example, when the population weight is increased, the total population of the created pre-EAs will be more homogeneous and closer to the maximum population threshold. Conversely, increasing the shape factor coefficient will result in more compact shapes.

Post-Processing: The quality of the input datasets primarily determines the quality of the outputs generated by the preEA tool. Due to limitations in the input datasets and the imposition of various constraints within the tool, some of the generated pre-EA outputs require assessment and modification. This is the main reason the outputs are referred to as "pre-EAs" rather than final EAs. For example, pre-EAs with negligible populations or geographic areas were automatically merged with neighbouring units using the "Eliminate Small Area" tool in the preEA package. Manual modifications were applied to pre-EAs covering large geographic areas with high populations.

3.3 Further Manual Adjustments

The national sampling frame created using the automatic approach requires additional modifications for several reasons. First, adjustments are necessary in areas where the existing datasets were insufficient to create smaller EAs. Second, the output contains some polygons with positive populations but no actual settlements, which is a limitation of the gridded population and settlement data, not the preEA tool. PreEAs with positive populations that fall within non-residential areas, such as offices, buildings under construction, recreational centres, factories, manufacturing zones, and agricultural lands, are manually adjusted.

3.4 Field support and guidance

This work also developed detailed automatic field maps and offline maps to support ground data collection and provide guidance. Several informative geospatial layers and techniques were employed in the process. For further details, please refer to Appendix B.

4 Results

4.1 Urban and Rural Classification

This paper presents Armenia's first accessible and usable digitized urban and rural boundaries (Figure 6). Based on the generated boundaries and the 2020 WorldPop gridded population data, urban areas constitute 20% of the land area, while rural areas cover the remaining 80%. In terms of population, over 60% of Armenians reside in urban areas, with less than 40% living in rural areas.

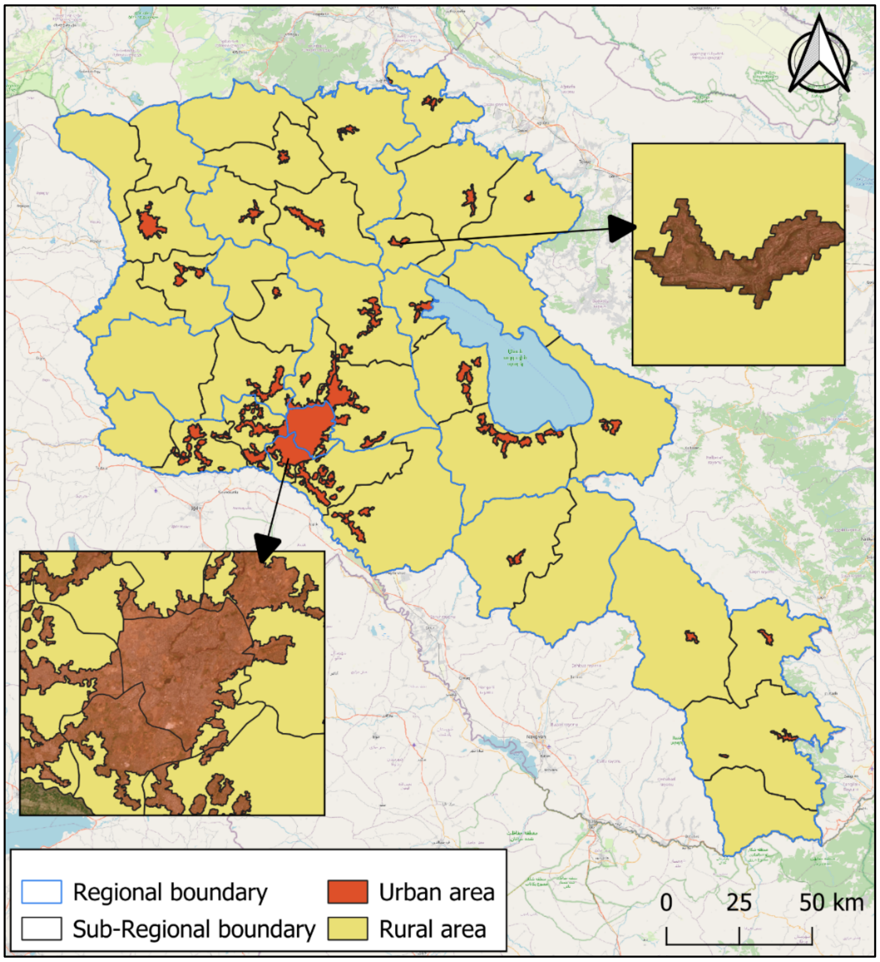

Figure 6: Generated urban and rural areas in Armenia. Basemap: OSM Standard.

4.2 A comparative analysis of urban and rural population: 2011 census vs. WorldPop gridded data

Grid-level population estimates are derived from WorldPop data, which can be aggregated to align with any specified geographic boundaries. In contrast, the 2011 Census data from the Committee of the Republic of Armenia (ArmStat) provides population counts, though with limited spatial detail, focusing primarily on marz and settlement types. Despite the restricted spatial information in the census data, the population estimates from WorldPop have been compared to the actual population count at the aggregate level to assess the accuracy of these estimates and the generated urban and rural classification. The population distribution at both the marz and urban/rural levels is highly comparable. The correlation coefficient between populations from the two data sources is 0.99 (SE: 0.05, p-value: 0.00) for marz-level populations and 1.0 for urban/rural populations. Figure 7 compares the "urban" population at the marz level, while Figure 8 compares the "total" population at the urban/rural level. These comparisons suggest that the population estimates from WorldPop are consistent with actual population data, at least at the marz and urban/rural levels. More importantly, it indicates that the automatic creation of urban and rural boundaries has produced an accurate classification.

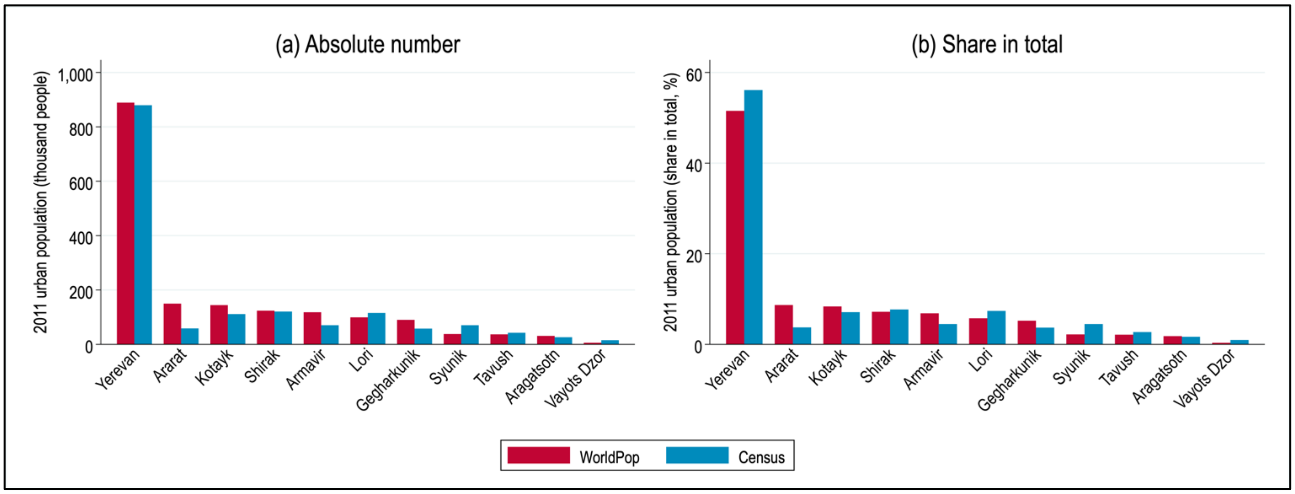

Figure 7: Urban population in *marzes*, 2011

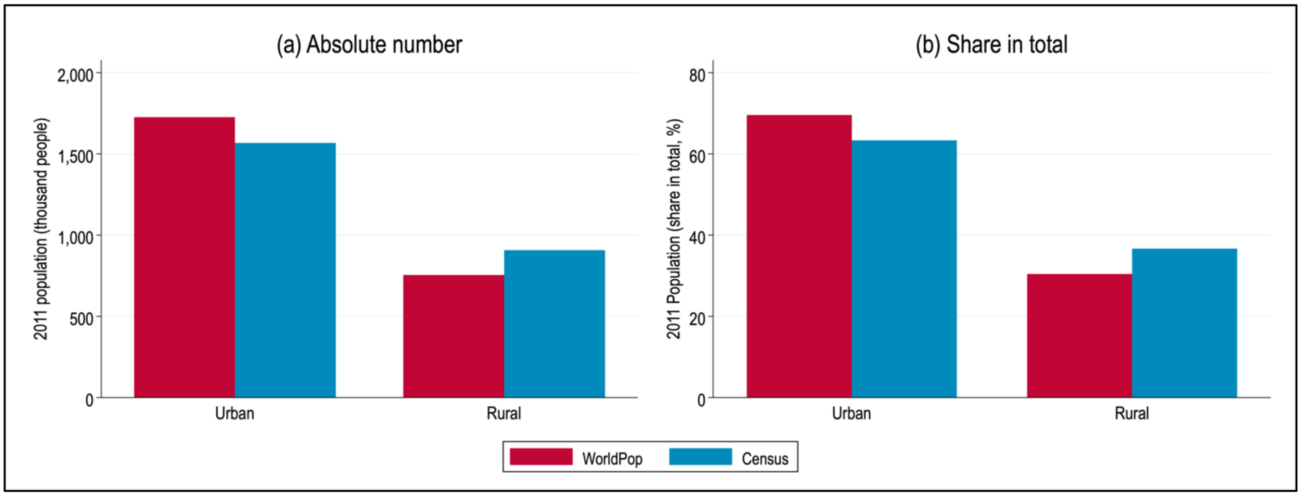

Figure 8: Urban and rural population, 2011

4.3 Armenia's National Sampling Frame

Figure 9 illustrates the pre-Enumeration Areas (pre-EAs) generated in this study using the method outlined in Section 3. The map demonstrates that the pre-EAs cover the entire territory of Armenia, with boundaries drawn accurately, free from geometric errors. In the initial stage of automatic pre-EA production, the pre-EA tool generated 130,378 building blocks (Figure 9a). Following the merging process, 7,413 pre-EAs were delineated across Armenia, with 3,813 in urban areas and 3,600 in rural areas (Figure 9b). After manual adjustments, approximately 60% of the pre-EAs (4,354) have a population greater than zero, most of which fall within the estimated range of 100 to 1,000 people. The remaining 3,059 pre-EAs are classified as unsettled, meaning their population is zero, as explained in Section 3.

While the map encompasses the entire country, about 41% of the area depicted would not be considered in the sampling designs, as the probability of selection for empty Primary Sampling Units (PSUs) is zero. Nonetheless, all individuals are accounted for in the population estimates.

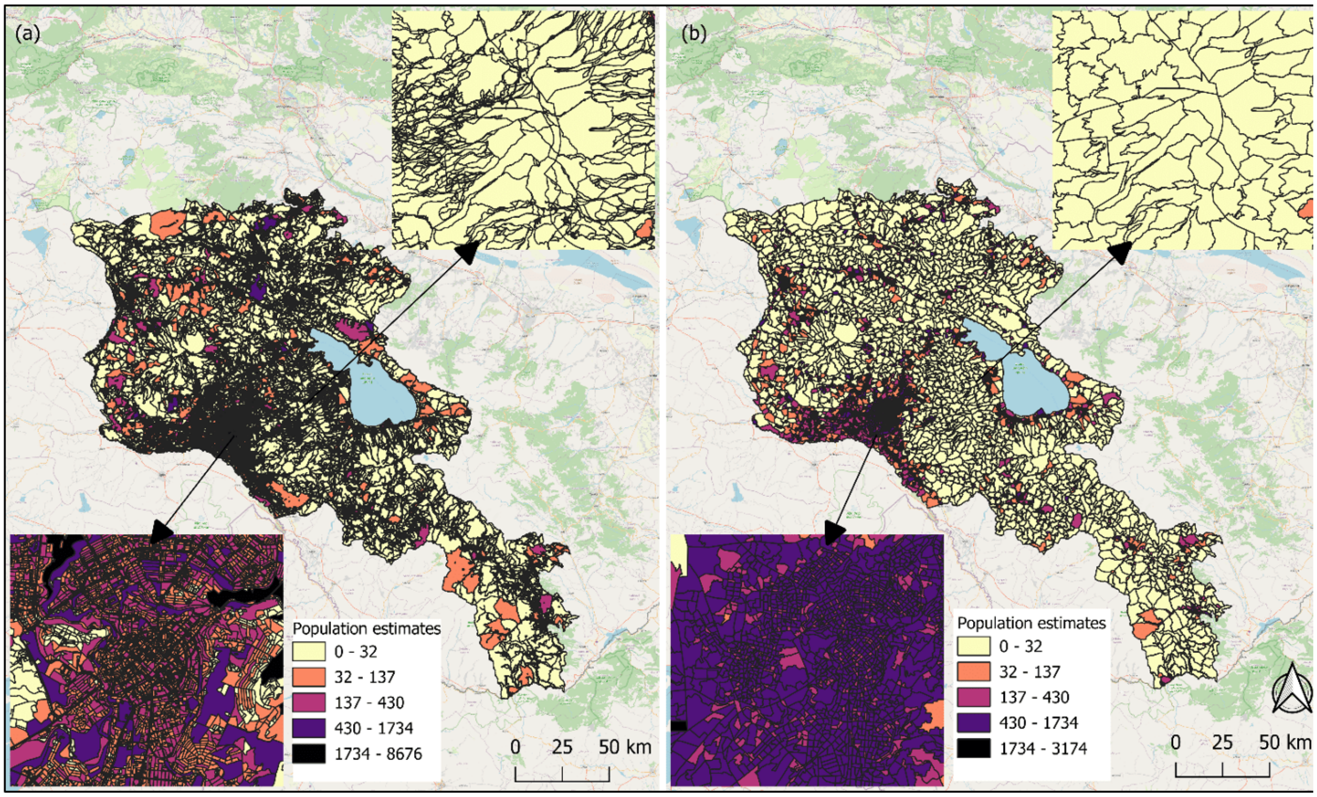

Figure 9: PreEA tool outputs

Notes: Panel (a) depicts building blocks before merging, while panel (b) shows pre-EA outputs after merging. Basemap: OSM Standard.

Next, the Primary Sampling Unit (PSU) size is characterized to assess the sampling frame and analyze the population distribution across pre-EAs, with particular emphasis on those pre-EAs with a positive population. Figure 10 displays the distribution of the 2022 population estimate across the pre-EAs in Armenia, which has been estimated based on two datasets. The 2020 population data is first sourced from WorldPop, which was projected based on the 2011 Census (Bondarenko et al., 2020). This population at the pre-EA level has then been re-scaled to match the 2020 population data from the ArmStat at the strata level that we described in Section 2. Then the preEA-level rescaled 2020 population data has been further projected to 2022 using strata-level population growth calculated from the ArmStat's population data over time.

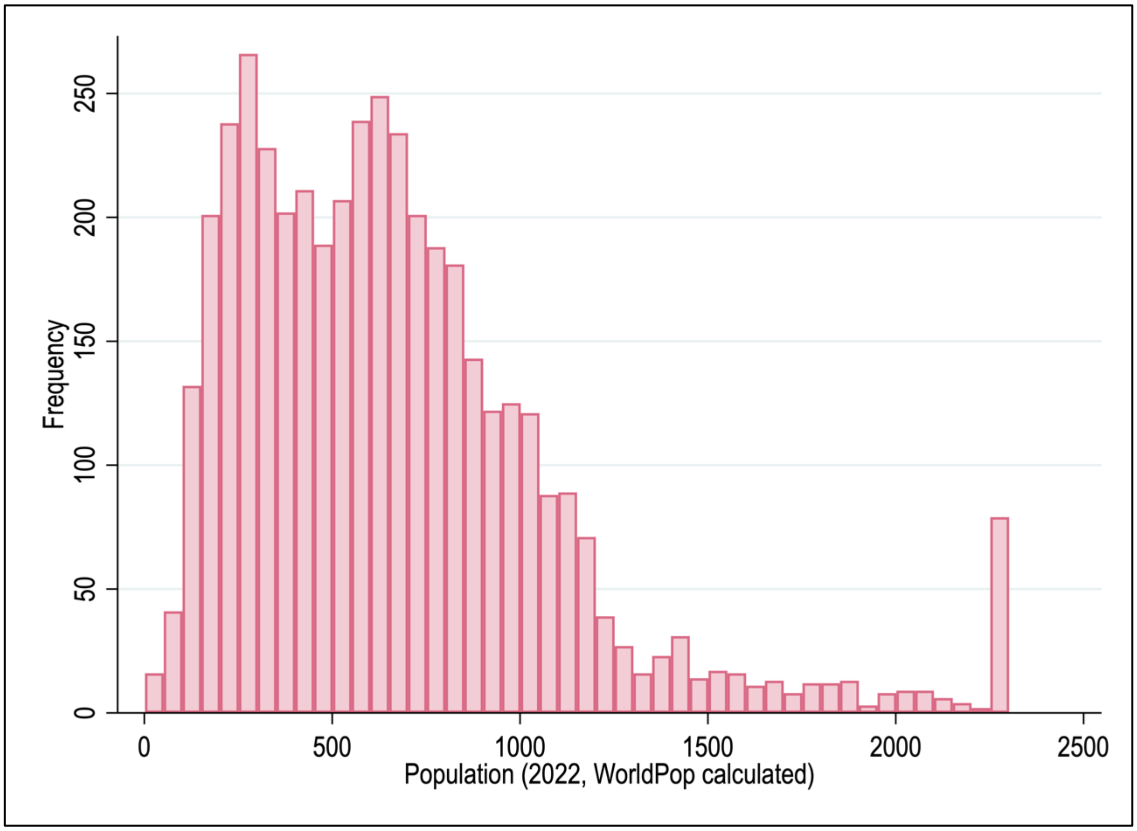

Figure 10: Distribution of population across pre-EAs in Armenia

Notes: The figure presents the distribution of population across pre-census enumeration areas (pre-EAs) in 2022. The extreme values of the population have been winsorized at the 1st and 99th percentiles to account for the potential outliers, i.e., we set the low (high) values at the 1st percentile (99th percentile).

The population distribution across the pre-EAs is fairly normal: 90% of the PSUs have a population below 1,156, while the remaining 10% have populations ranging from 1,156 to 2,274. These values were winsorized[2] at the 1st and 99th percentiles to minimize the impact of potential outliers. The population of the pre-EA is not winsorized in the sample frame, and it is our suggestion to adjust those extreme values as the population might have been overestimated in those areas. Users of this sampling frame can make their own judgments regarding these values. Despite the presence of some outliers at the upper end of the distribution, the population across pre-EAs is generally more evenly distributed than the adult population across electoral precincts, which displays U-shaped patterns, as shown in Figure 1.

[2] In statistics, winsorizing replaces extreme values (outliers) in a dataset with less extreme values, typically the values at the specified percentiles, to reduce the impact of those outliers in statistical analysis.

Figure 11 presents examples of pre-EAs with positive populations located within non-residential areas. As these pre-EAs do not contain residents, their population estimates are adjusted by setting them to zero. Prior to this adjustment, 1,601 pre-EAs already had a population of zero, and the population of 1,458 additional pre-EAs was revised to zero. This results in a total of 3,059 pre-EAs being classified as unsettled areas. Following the adjustment, the population estimates for the remaining pre-EAs with positive populations were further refined by applying a strata-level factor to align them with the strata-level population counts.

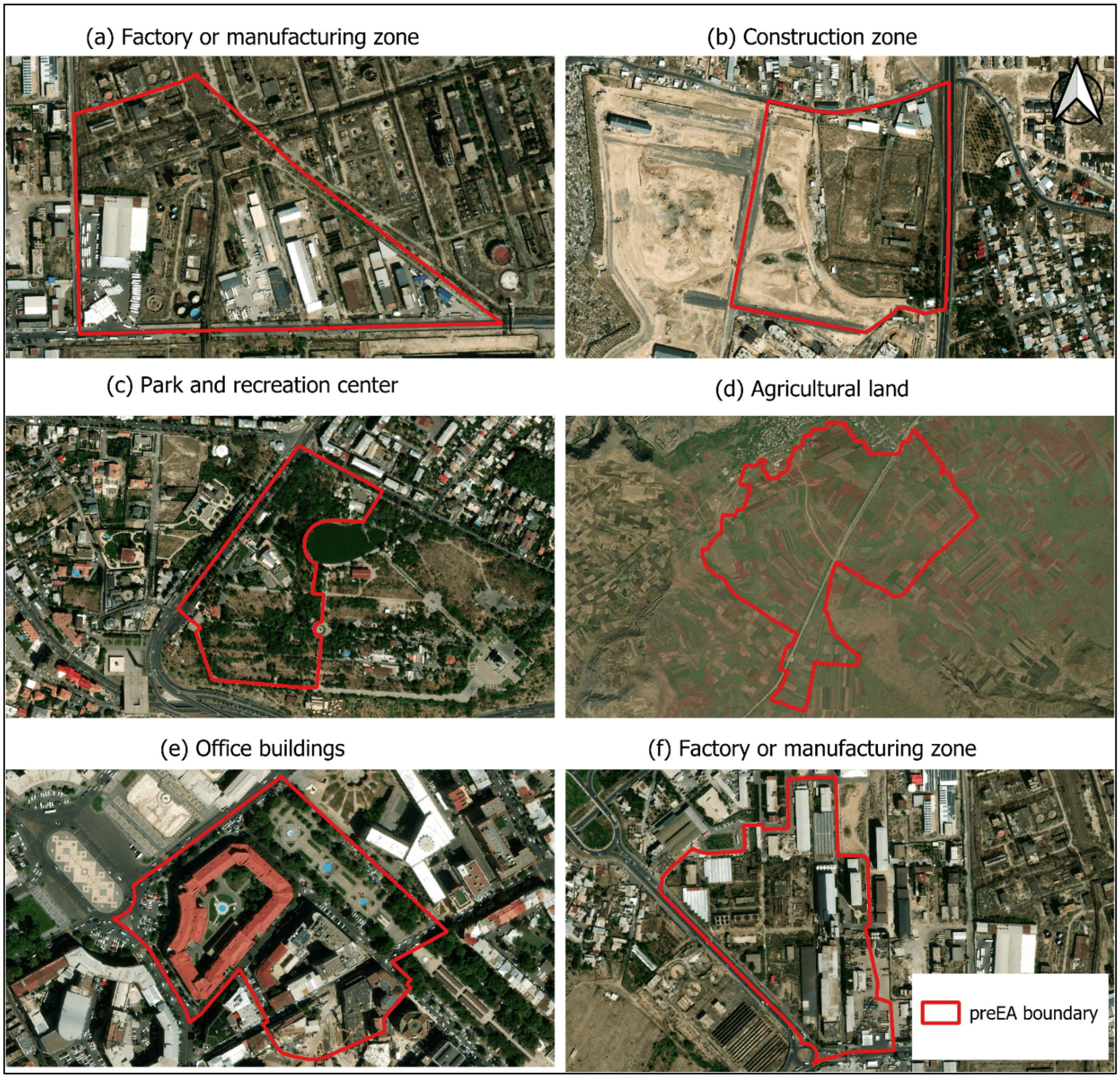
(a) Factory or manufacturing zone
(b) Construction zone
(c) Park and recreation center
(d) Agricultural land
(e) Office buildings
(f) Factory or manufacturing zone
preEA boundary

Figure 11: Examples of pre-EAs located in non-residential areas yet have a population greater than zero. Basemap: ESRI Satellite Imagery.

Certain international guidelines must be adhered to when developing pre-EA boundaries, as illustrated in Figure 12. Within the designated administrative boundaries, the pre-EA boundaries must be nested (Figure 12a). Features such as rivers and major roads, which are considered uncrossable, must be respected when defining pre-EA outlines (Figure 12b). Additionally, the pre-EA boundaries should align with visible ground features, such as roads and infrastructure. Figure 12c provides examples of pre-EA boundaries in urban areas, while Figure 12d shows the pre-EA boundaries in rural areas. These figures demonstrate the extent to which the pre-EA outlines align with discernible ground features, highlighting their accuracy in reflecting the physical landscape.

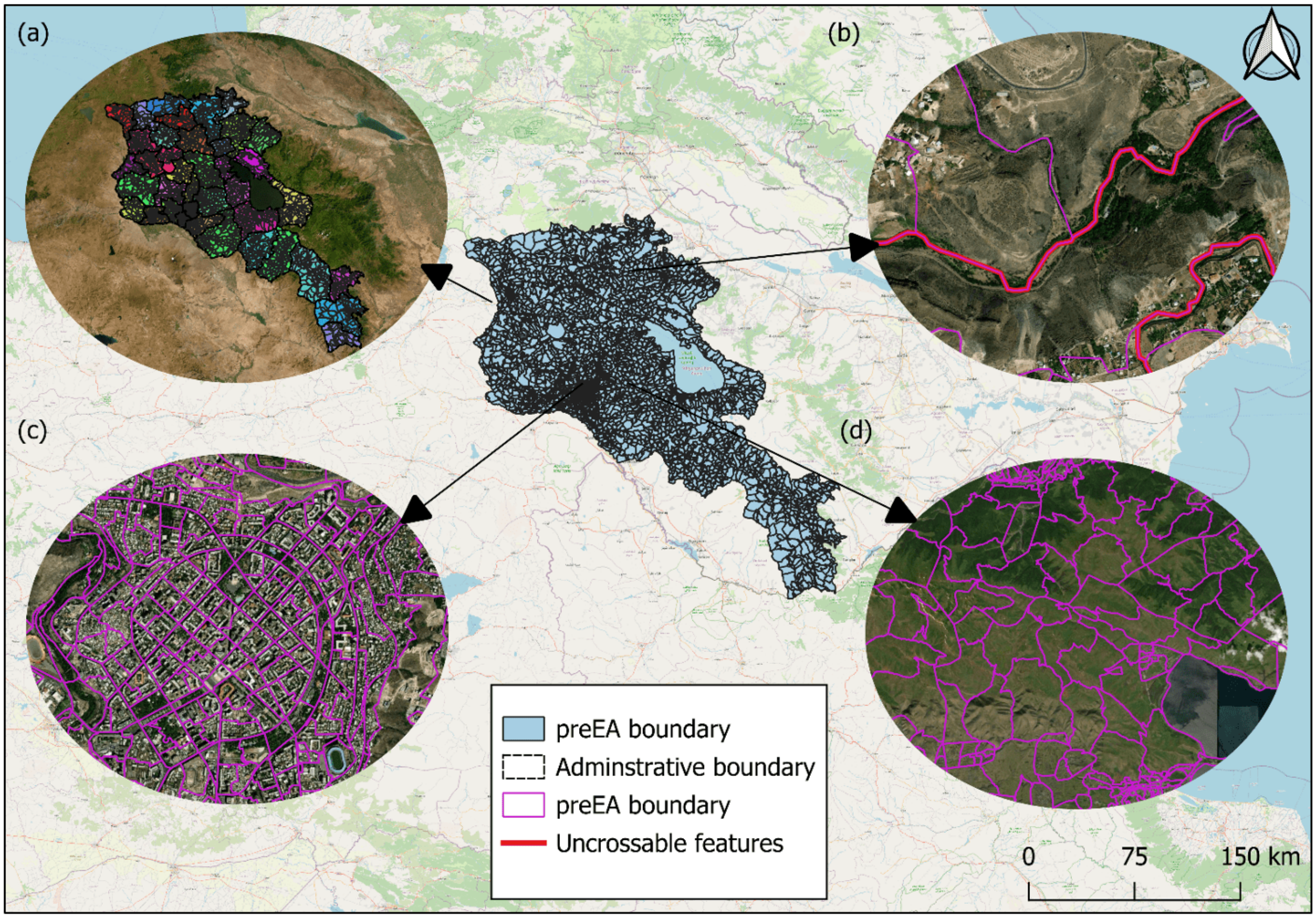

Figure 12: The outline of pre-EA boundaries. Basemap: OSM Standard and ESRI Satellite Imagery.

4.4 The First Application: Listening to Armenia

This section presents an application of the proposed national sampling frame based on pre-EAs in Armenia, specifically describing the sampling design of the World Bank Group's "Listening to Armenia" survey (L2Arm). The application of the proposed sampling frame in the L2Arm survey has proven to be a successful approach. It allowed for a nationally representative sample that met the objectives of capturing urban and rural distinctions across all regions of Armenia. The innovative use of pre-EAs optimizes resource allocation and ensures the feasibility of the survey within time and budget constraints.

4.4.1 Overview of the Survey

Listening to South Caucasus (L2SC) is an ongoing project and an expansion of a collaborative effort that has been conducted in multiple countries in the Europe and Central Asian region.[3] This initiative aims to comprehensively monitor the views and well-being of a representative group of people as the government introduces social and economic reforms that affect every business and citizen. By reflecting on the experience of this group over the years, the study provides an up-to-date understanding of how policies reflect on people's daily lives. The study comprises a nationally representative baseline survey and a high-frequency panel survey of a subset of the baseline participant households. The information collected through the L2SC initiative informs reform efforts directly by raising the profile of citizens' views and enabling in-depth economic analysis. While the L2SC survey covers Armenia and Georgia, this paper focuses on the baseline survey in Armenia—Listening to Armenia (L2Arm)—where the new national sampling frame based on pre-EAs has been proposed.

4.4.2 Sampling Design

The sampling design optimizes the spatial allocation of the household sample to provide valid representativeness at the national level for both urban and rural areas. A two-stage stratified cluster sampling design is employed to select participating households, ensuring a balanced sample distribution across regions and accounting for differences between urban and rural areas, survey budgets, and discrepancies in population estimates. The L2Arm survey's implementation highlighted the robustness of the sampling frame, as it successfully captured the population distribution across diverse geographic and demographic strata. The use of probability proportional to size (PPS) sampling ensured that the selection process was equitable and aligned with population estimates, further validating the practicality of the proposed approach.

In the first stage, a certain number of primary sampling units (PSUs) will be selected in each urban and rural stratum (urban and rural areas within each administrative region). In the second stage, the ultimate sampling units or the secondary sampling units (SSUs)—households in the case of L2SC—are randomly selected within each PSU. The survey is then implemented among the selected households. Given that our focus in this paper is on the first stage of the two-stage procedure, the sampling frame used for the survey in the first stage is highlighted.

Sampling Frame: As mentioned, the sampling frame is based on pre-census enumeration areas (pre-EAs) providing the most accurate information on the geographic distribution of the population across Armenia. This allows for the most precise formulation of a sample design.

Sample Size: The objective of any sample design is to achieve the highest precision in indicators of interest given survey parameters. The sampling design aims to efficiently allocate the given PSUs across strata and the households across PSUs. For L2Arm, 400 PSUs are allocated across strata proportional to the population (i.e., implicit allocation) with some adjustments. Ten households are targeted within each PSU. Table 5 presents the proposed baseline sample allocation.

[3] The in-progress L2SC survey is not publicly available yet. However, the survey is available for other Europe and Central Asian (ECA) countries, including Uzbekistan, Kazakhstan, the Kyrgyz Republic, Tajikistan, and Ukraine.

In the first stage of the two-stage stratified cluster sampling design, PSUs in each stratum are randomly selected using systematic random sampling with probability proportional to size (PPS), size being the estimated population of the pre-EA. This method assigns each PSU's likelihood of selection based on the PSU's size within the stratum. Population size, rather than the number of households, is used due to a lack of data on the number of households at the PSU level in the sampling frame. Thus, each PSU's likelihood of selection corresponds to the percentage of the stratum population residing in the PSU. In the second stage, a set number of households are randomly selected from each chosen PSU.

Table 5: Baseline sample design based on proportional sample allocation

Stratum	Baseline			
	2022 Population Estimate	Allocated Number of PSUs	Target Number of HHs (PSU size)	Number of HHs
Yerevan	1,098,866	147	10	1470
Aragatsotn - Urban	26,738	4	10	40
Aragatsotn - Rural	98,949	13	10	130
Ararat - Urban	72,294	10	10	100
Ararat - Rural	186,983	25	10	250
Armavir - Urban	82,953	11	10	110
Armavir - Rural	183,703	25	10	250
Gegharkunik - Urban	65,902	9	10	90
Gegharkunik - Rural	162,809	22	10	220
Kotayk - Urban	137,493	18	10	180
Kotayk - Rural	116,364	16	10	160
Lori - Urban	124,050	17	10	170
Lori - Rural	87,532	12	10	120
Shirak - Urban	133,620	18	10	180
Shirak - Rural	96,856	13	10	130
Syunik - Urban	90,205	12	10	120
Syunik - Rural	44,350	6	10	60
Tavush - Urban	49,859	7	10	70
Tavush - Rural	69,943	9	10	90
Vayots Dzor - Urban	16,160	2	10	20
Vayots Dzor - Rural	31,501	4	10	40
Armenia	2,977,130	400		4000

Notes: The population estimates aggregated at the stratum level are by the end of 2022 and match the population statistics from the Statistical Committee of the Republic of Armenia.

5 Discussions

The national sampling frame based on pre-EAs offers several advantages that are not provided by existing and accessible potential sampling frames. However, it may also present practical challenges and methodological limitations. This section discusses the additional benefits and potential concerns associated with this approach and proposes solutions. These solutions have been successfully tested in other developing countries where pre-EAs have been implemented, such as Somalia (Qader et al., 2021) and the Democratic Republic of the Congo (Qader et al., 2023).

Population Estimates as a Measure of EA Size: The primary challenge of using the proposed national sampling frame for household surveys is that the size of the pre-EA is based on population estimates derived from gridded population data. These estimates may differ from the actual population, potentially introducing bias in the probability of preEAs being selected during the first stage of the two-stage design. Although this study does not address the validation of the gridded population estimates, it is important to clarify the limitations of the data. The automatic creation of a national sampling frame requires granular population information to ensure that the resulting sampling units are manageable. However, this level of granularity is not available in the existing census data in Armenia. As a result, gridded population data is utilized. In developing countries, several gridded population datasets with varying spatial resolutions are accessible, including Gridded Population of the World (GPWv4) (CIESIN, 2016), WorldPop (Bondarenko et al., 2020), High-Resolution Settlement Layer (HRSL) (Facebook and CIESIN, 2016), Demobase Population datasets (Azar et al., 2013), Global Human Settlement Population Grid (GHS-POP) (JRC, 2015), Global Rural-Urban Mapping Project (GRUMP) (CIESIN, 2011), and LandScan (Dobson et al., 2000). The accuracy and quality of gridded population data are primarily influenced by the quality of the input data model, which includes census data, satellite-derived covariates, and the statistical model used. For this study, the WorldPop-constrained gridded population data for Armenia from 2020 was used to create the national sampling frame, as it was the most recent available with reasonable spatial resolution. This implies that there may be notable differences between the population size and distribution in 2020 and the present day. However, since users can update the national sampling frame's population based on their desired data sources, such as a population registry, this discrepancy should not be a major concern.

Our findings indicate that the total populations of pre-EAs vary, ranging from zero to a specific population size. In the pre-EA tool, users can define various constraints, with the maximum population size and geographic area being the two primary hard constraints. These maximum thresholds may vary depending on the objectives of the work or the specific country context. The main purpose of establishing maximum thresholds for both population and area is to balance these limits and prevent the creation of unmanageable preEAs in areas with sparse populations. Once one of these constraints is met, aggregation ceases during the merging process. In uninhabited areas (as indicated by gridded population data), the size of preEAs is determined solely by the maximum geographic area; if this threshold is reached, aggregation stops. Consequently, several pre-EAs with zero or low population values may be created. This issue can be addressed in the tool by removing the geographic area constraint, but doing so may result in the creation of excessively large pre-EAs that could be difficult to enumerate, particularly in rural areas. The primary benefit of considering geographic constraints is that it helps avoid including uninhabited areas in sampling surveys, leading to significant time and cost savings. However, as the method primarily relies on gridded population data, there is a risk that some inhabited areas may be overlooked if the data is inaccurate or unreliable. The severity of bias due to using population estimates as PSU size depends on the size of the discrepancy between the actual population and population estimates and whether the difference is systematic.

This paper presents the first accessible and usable urban and rural classification for Armenia, contributing to the development of a national sampling frame. Currently, there is no available digital urban and rural boundary that can be compared with the boundaries generated in this study. However, we compared urban and rural population estimates between the boundaries we generated and those from the 2011 census. While there is a strong correlation between the aggregated population estimates from the census and the gridded population estimates for urban and rural areas, as discussed in Section 5, the output may not fully reflect reality. This is primarily due to the GHSL SMOD's approach, which classifies the world into urban and rural categories using gridded population data and built-up areas derived from various data sources. The algorithms employed to generate the input data for both datasets, along with the satellite imagery used to extract the covariates, can introduce certain biases, affecting the accuracy of the classification. Furthermore, National Statistical Offices (NSOs) often use non-standardized approaches to classify urban and rural areas within their countries.

Another challenge associated with the inaccuracy of gridded population data is the occasional allocation of people in non-residential areas. This issue arises when the data fails to properly distinguish between residential and non-residential spaces, leading to an incorrect assignment of the population in preEAs that do not contain residents (Archila et al. 2020; Kuffer et al. 2022). Model estimates of gridded population data can be improved with a reliable approach, along with sufficient resources and data, to accurately identify non-residential buildings. However, this remains a significant challenge due to the complexity of the issues involved, such as the similarity of structures and the coexistence of residential and commercial tenants within the same building (Hu et al. 2015; Han et al. 2017; Chew et al. 2018). As a result, non-residential areas are not always excluded in the population predictions of various gridded population datasets. Consequently, some pre-EAs located within non-residential areas may still show positive population values.

It is important to note that when the pre-EAs were verified against high-resolution satellite imagery base maps from ESRI and Google, many of these validations were based on visual observation. Since the dates of the satellite imagery were not considered, these evaluations may not have been entirely objective. Therefore, without comprehensive validation on the ground, these assumptions cannot be fully verified.

Digitized Elements and Boundaries: Digitalized elements, both natural and man-made, are crucial for the automatic generation of the national sampling frame. In this study, the method leveraged the extensive digital line data from OpenStreetMap (OSM), which includes roads, railways, and waterways. However, the pre-EAs generated often exceeded the specified thresholds, such as population and geographic area, due to the poor quality and incomplete spatial coverage of the existing digitized boundaries. The main causes of this issue are (i) incomplete and (ii) disconnected lines. If certain natural and artificial features remain undigitized, further work—either manually or automatically—needs to be carried out. Lines should never be left open and should always be connected to other features whenever possible. This is because disconnected lines will not be polygonised during the polygonization process, which leads to the creation of larger, unmanageable preEAs. In addition to spatial coverage, the quality of OSM attribute data is essential for the automatic creation of the national sampling frame. Several line features, including major roads, rivers, and other barriers, were classified as uncrossable to improve the collection of ground data and enhance efficiency. The only source that can accurately determine the types of features on the ground is the attribute information. If the feature classification in the attribute table is incorrect, it may result in misclassification of uncrossable features, thereby impacting the accuracy of the national sampling frame. The quality, spatial coverage, and attribute information of OSM data may vary from one country to another (Haklay, 2010; Barrington-Leigh and Millard-Ball, 2017; Minaei, 2020; Gatea and Al-Bakri, 2023).

The semi-automatic approach creates pre-EAs based on digitized visible ground features, which are generally unlikely to intersect with buildings or other structures. However, there are instances where such intersections may occur. These intersections may be caused by administrative boundaries or poorly entered visible digital lines. Since administrative boundaries cannot be altered without consulting the relevant government agencies, users should be cautious when determining the reasons for the cutting of buildings and other structures. For example, Figure 13 illustrates a pre-EA output where the boundary cuts through buildings. This is due to the administrative boundary of the municipality, and as such, it cannot be modified.

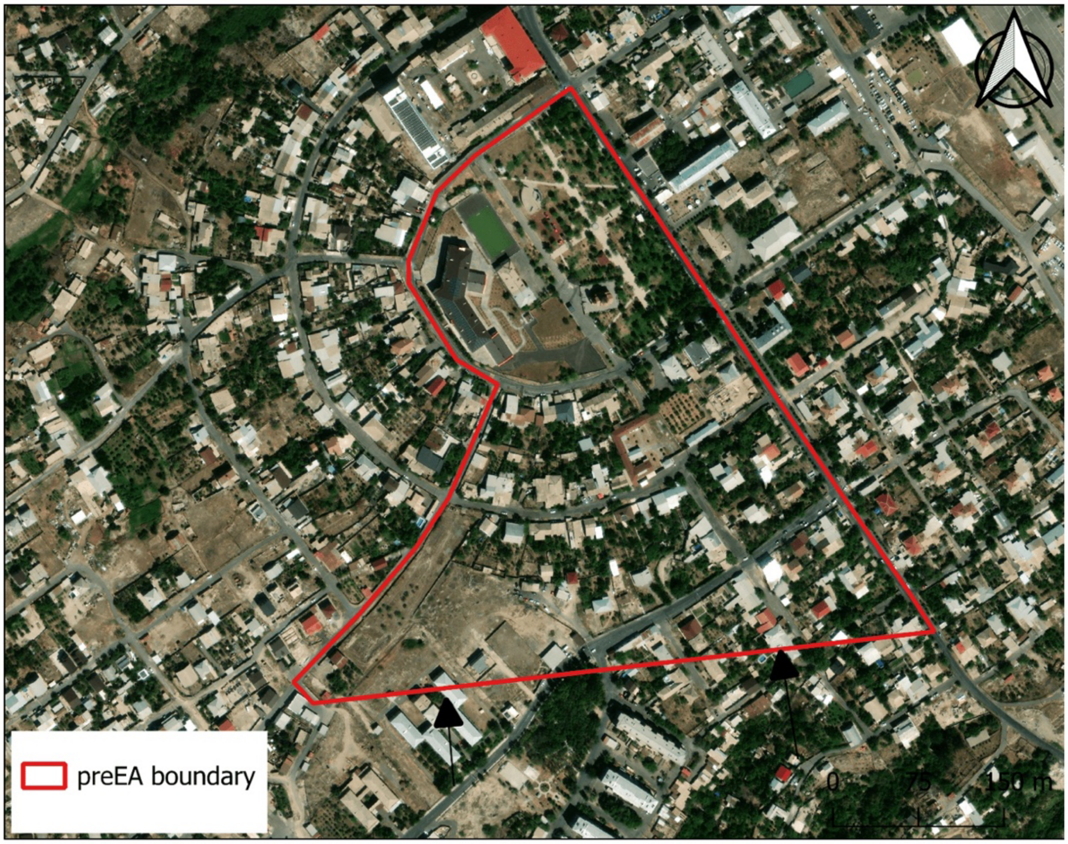

Figure 13: An example of pre-EA boundary-cutting buildings

This method has solely utilized publicly accessible natural and man-made features, and settlement boundaries, such as OSM and GHSL, for reproducibility and worldwide application. Nonetheless, several government agencies can provide input datasets such as roads and waterways with higher quality and greater geographic coverage. In addition, future studies could also investigate leveraging the more modern and comprehensive commercial road network (Strano et al., 2017). If inadequate spatial coverage is a major concern, an alternative approach would be to use "mapathons"—a coordinated mapping event—to enhance the current open-source data on roads and rivers before implementing this method.

One of the primary challenges in collecting high-quality surveys in Armenia was the absence of clearly defined boundaries for PSUs. To our knowledge, it remains uncertain whether the Statistical Committee of the Republic of Armenia (ArmStat) possesses a digital map of census enumeration areas. As a result, this paper presents the first accessible digitized national sampling frame for the country. The inclusion of PSU boundaries and other administrative units in our sampling frame provides several advantages. Notably, it helps prevent errors such as the inclusion of households outside the designated survey areas. If such errors occur non-randomly, they could significantly compromise the integrity of subsequent analyses based on the collected data. Furthermore, this type of error may be systematic, particularly for PSUs with larger areas and longer boundaries, where such mistakes are more likely to occur. Therefore, this feature plays a crucial role in ensuring robust quality control throughout the data collection process.

It is difficult to directly compare the resources and budget of our automated method with manual approaches, as many countries do not offer a detailed breakdown of the costs associated with various stages of census operations, especially the resources needed for manually digitizing national census enumeration areas. For example, the 2010 census mapping effort in Zambia was projected to cost approximately US $7 million and take nearly two years to complete (United Nations Secretariat, 2007). If the pre-census sampling frame in Armenia had been manually digitized, significant financial resources would have been required to extensively train a team of cartographers on how to digitize all the units using high-resolution satellite imagery. Additionally, the entire pre-enumeration area would need to be manually digitized in accordance with strict requirements, necessitating considerable effort to ensure quality control and correct geometric errors, given the hand-drawn nature of the process. This approach would have been both time- and resource-intensive. In contrast, the automated creation of Armenia's pre-census sampling frame was completed in under three months, including the manual corrections needed due to the lack of spatial input data and feedback was received by the local experts, all carried out by a single specialist. The significant savings in labour, time, and costs from using an automated method can be reinvested into other aspects of national surveys and census preparation, enhancing overall efficiency.

Despite its limitations, the method was successfully implemented, resulting in the creation of a national sampling frame for Armenia. From financial, time, and technological perspectives, this approach outperformed conventional manual techniques. Historically, manually delineating a nationwide sampling frame required years of work and substantial financial resources. Moreover, the manual method is susceptible to various geometric issues, such as gaps, overlaps, pockets, and disjunctions, due to the inherent limitations of human error. These geometric inconsistencies could introduce bias into the sampling frame and, consequently, the data collected. In contrast, the automatic method eliminates these geometric problems, ensuring greater accuracy. Furthermore, the automatic approach offers several advantages over the gridded population sampling frame, which has been directly used as a sampling frame in various studies (Thomson et al., 2017; Cajka et al., 2018; Qader et al., 2020). The key difference between our approach and the gridded population frame lies in the design of the sampling units. In gridded population methods, buildings and other structures are often truncated because the grid's boundaries do not align with visible features on the ground. In contrast, the preEA tool generates pre-enumeration area boundaries that follow observable, natural features such as rivers and roads, providing a more accurate and relevant sampling frame.

Potential Applications of the Method in Different Countries: The absence of a national sampling frame presents significant challenges during the implementation stage of many household surveys. In some countries, an up-to-date and digitized national sampling frame may not be available. While such a frame may exist in other countries, National Statistical Offices (NSOs) may be unwilling to grant access to international agencies such as the World Bank. In some cases, the sampling frame relies on census enumeration areas, which, due to their large spatial units, can lead to substantial costs when conducting the second and third stages of sampling to achieve the required household size. Additionally, if the sample selection is based on census enumeration areas, the household listing for the selected sampling units requires considerable time and resources due to their extensive spatial coverage and large population totals. Our proposed method and strategy offer a potential solution for creating a new national sampling frame in the event of these challenges arising in future surveys in other countries.

6 Conclusion

This paper introduces a new national sampling frame for the Republic of Armenia, serving as a model for developing nations with limited access to functional sampling frames for representative household surveys and potentially future censuses. Specifically, it presents an innovative method for the automatic delineation of pre-census enumeration areas (pre-EAs), which offers several advantages over traditional sources of sampling frames such as those based on outdated census enumeration areas, census settlements, electoral precincts, and traditional gridded sampling techniques.

The national sampling frame developed in this paper divides Armenia into approximately 7,500 pre-EAs, the majority of which have population estimates greater than zero. These estimates, which are recent and relatively homogeneous, range mostly between 100 and 1,000 people. The digitized enumeration areas with clearly defined boundaries facilitate the household selection process by ensuring that households outside of the selected PSUs are not included.

This paper makes several methodological and practical contributions to the survey sampling literature and to organizations that collect and utilize representative surveys, such as researchers and policymakers. First, it expands the application of the semi-automatic approach for creating national sampling frames by generating Armenia's first digitized frame based on pre-EAs, offering an alternative to traditional methods of delineating national sampling frames. Our analysis highlights the applicability of pre-EAs for other countries facing similar challenges in developing sampling frames. Second, the national sampling frame contributes to survey implementers and users of household surveys in Armenia by providing a standardized and decentralized framework. Third, the paper systematically evaluates the existing sampling frames in Armenia, comparing their strengths and limitations to the proposed frame. This comparison suggests that our frame complements existing sampling frames and can serve as a viable alternative.

In conclusion, the paper acknowledges some limitations and outlines directions for future research. While the proposed national sampling frame addresses a common challenge in the first stage of two-stage sampling designs, solutions for challenges encountered in the second stage, such as household listing strategies, are beyond the scope of this paper. Future research could explore innovative approaches to household listing, particularly when utilizing the sampling frame introduced here.

References

Archila Bustos, M. F., Hall, O., Niedomysl, T., and Ernstson, U. (2020). "A pixel level evaluation of five multitemporal global gridded population datasets: a case study in Sweden, 1990–2015." *Population and Environment,* 42(2), 255-277. https://doi.org/10.1007/s11111-020-00360-8.

ARMSTAT. (2025). "The Results of 2022 Population Census of RA." https://armstat.am/en/?nid=944 (Accessed: 08/04/2025)

ARMSTAT. (2022). "The Necessity of Population Census Conduction and Ensuring Its Legal Basis." https://www.armstat.am/file/article/introduction.pdf.

ARMSTAT. (2024). "The Results of 2011 Population Census of the Republic of Armenia (Indicators of the Republic of Armenia)." https://armstat.am/en/?nid=532.

Azar, Derek, Ryan Engstrom, Jordan Graesser, and Joshua Comenetz. (2013). "Generation of Fine-Scale Population Layers using Multi-Resolution Satellite Imagery and Geospatial Data." *Remote Sensing of Environment*, 130: 219–232. https://doi.org/10.1016/j.rse.2012.11.022.

Barrington-Leigh, Christopher, and Adam Millard-Ball. (2017). "The World's User-Generated Road Map is More than 80% Complete." *PLoS ONE* 12(8): e0180698. https://doi.org/10.1371/journal.pone.0180698.

Bondarenko, Maksym, David Kerr, Alessandro Sorichetta, and Andrew Tatem. (2020). "Census/Projection-Disaggregated Gridded Population Datasets for 189 Countries in 2020 using Built-Settlement Growth Model (BSGM) Outputs." University of Southampton https://dx.doi.org/10.5258/SOTON/WP00684 [Dataset].

Brown, J. A., B. L. Robertson, and Trent McDonald. (2015). "Spatially Balanced Sampling: Application to Environmental Surveys." *Procedia Environmental Sciences*, 27: 6-9. https://doi.org/10.1016/j.proenv.2015.07.108.

Cajka, James, Safaa Amer, Jamie Ridenhour, and Justine Allpress. (2018). "Geo-Sampling in Developing Nations." *International Journal of Social Research Methodology*, 21(6): 729–746. https://dx.doi.org/10.1080/13645579.2018.1484989.

Center for International Earth Science Information Network - CIESIN - Columbia University, International Food Policy Research Institute - IFPRI, The World Bank, and Centro Internacional de Agricultura Tropical - CIAT. (2011). "Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Population Count Grid." Palisades, New York: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4VT1Q1H.

Center for International Earth Science Information Network - CIESIN - Columbia University. (2016). "Gridded Population of the World, Version 4 (GPWv4): Administrative Unit Center Points with Population Estimates." Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4F47M2C.

Central Bureau of Statistics of Nepal. (1996). "Nepal Living Standard Survey – 1995: Basic Information Document." https://microdata.worldbank.org/index.php/catalog/2301/download/34437.

Chew, R. F., Amer, S., Jones, K., Unangst, J., Cajka, J., Allpress, J., and Bruhn, M. (2018). "Residential Scene Classification for Gridded Population Sampling in Developing Countries using Deep Convolutional Neural Networks on Satellite Imagery." *International Journal of Health Geographics*, 17(1), 12. https://doi.org/10.1186/s12942-018-0132-1.

Darin, E., Dicko, A. H., Galal, H., Jimenez, R. M., Park, H., Tatem, A. J., and Qader, S. (2024). "Mapping Refugee Populations at High Resolution by Unlocking Humanitarian Administrative Data." *Journal of International Humanitarian Action*, 9(1), 14. https://doi.org/10.1186/s41018-024-00157-6.

Dobson, Jerome E., Edward A. Bright, Phillip R. Coleman, Richard C. Durfee, and Brian A. Worley. (2000). "LandScan: A Global Population Database for Estimating Populations at Risk." *Photogrammetric Engineering and Remote Sensing*, 66(7): 849–857. https://www.asprs.org/wp-content/uploads/pers/2000journal/july/2000_jul_849-857.pdf.

EUROSTAT. (2021). "Applying the Degree of Urbanisation: A Methodological Manual to Define Cities, Towns and Rural Areas for International Comparisons." Publications Office of the European Union. https://dx.doi.org/10.2785/706535.

Facebook Connectivity Lab and Center for International Earth Science Information Network - CIESIN - Columbia University. (2016). "High Resolution Settlement Layer (HRSL)." Source imagery for HRSL © 2016 DigitalGlobe. https://www.ciesin.columbia.edu/data/hrsl/.

Gale, R. P., Muradyan, A., Danelyan, S., Manukyan, N., Babak, M. V., Arakelyan, S., Tamamyan, G., and Arakelyan, J. (2023). "The Humanitarian Crisis in Nagorno-Karabakh." *The Lancet*, 402(10410), 1324-1325. https://doi.org/10.1016/S0140-6736(23)02034-2.

Gatea, Z. K., and Maythm, A. B. (2023). "Measuring the Attribute Accuracy and Completeness for the OpenStreetMap Roads Networks for Two Regions in Iraq." *Journal of Engineering*, 29(05), 156-168. https://doi.org/10.31026/j.eng.2023.05.12.

Geofabrik. (2025). "Latest OSM for Armenia." Available from: https://download.geofabrik.de/asia/armenia.html. (Accessed: 08/04/2025).

Global Human Settlement Layer. (2025). https://human-settlement.emergency.copernicus.eu/ghs_pop.php. (Accessed: 08/04/2025).

GRID3. (2025). "High Resolution Population Estimates." Available from: https://grid3.org/solution/high-resolution-population-estimates. (Accessed: 08/04/2025).

Grosh, Margaret E., and Paul Glewwe. (1995). "A Guide to Living Standards Measurement Study Surveys and their Data Sets." LSMS Working Paper No. 120. https://documents1.worldbank.org/curated/ar/270551468764720584/pdf/multi-page.pdf.

Haklay, M. (2010). "How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets." *Environment and Planning B: Planning and Design*, 37(4), 682-703. https://doi.org/10.1068/b35097.

Han, X., Zhong, Y., Cao, L., and Zhang, L. (2017). "Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote Sensing Image Scene Classification." *Remote Sensing*, 9(8), 848. https://www.mdpi.com/2072-4292/9/8/848.

Hansen, Henning S., and Lise Schrøder. (2019). "The Societal Benefits of Open Government Data with Particular Emphasis on Geospatial Information." In *Electronic Government and the Information Systems Perspective: 8th International Conference, EGOVIS 2019, Linz, Austria, August 26–29, 2019, Proceedings 8*. eds. by A. Kő, E. Francesconi, G. Anderst-Kotsis, A. Tjoa, and I. Khalil: Springer International Publishing, 31-44. https://doi.org/10.1007/978-3-030-27523-5_3.

HDX. (2025). "Armenia - Subnational Administrative Boundaries." Available at: https://data.humdata.org/dataset/cod-ab-arm. (Accessed: 08/04/2025).

Himelein, K., Eckman, S., and Murray, S. (2014). "Sampling Nomads: A New Technique for Remote, Hard-to-Reach, and Mobile Populations." *Journal of Official Statistics*, 30(2), 191-213. https://doi.org/10.2478/jos-2014-0013.

Howell, C. R., Su, W., Nassel, A. F., Agne, A. A., and Cherrington, A. L. (2020). "Area Based Stratified Random Sampling using Geospatial Technology in a Community-Based Survey." *BMC Public Health*, 20(1), 1678. https://doi.org/10.1186/s12889-020-09793-0.

Hu, F., Xia, G.-S., Hu, J., and Zhang, L. (2015). "Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery." *Remote Sensing*, 7(11), 14680-14707. https://www.mdpi.com/2072-4292/7/11/14680.

International Union for the Scientific Study of Population Toggle menu. (2025). "The Gridded Population of the World (version 4)." Available from: https://iussp.org/en/gridded-population-world-version-4-released. (Accessed: 08/04/2025).

Joint Research Centre (JRC). (2015). "GHS-POP R2015A - GHS population grid, derived from GPW4, multitemporal (1975, 1990, 2000, 2015) - OBSOLETE RELEASE." European Commission, Joint Research Centre (JRC). https://data.europa.eu/89h/jrc-ghsl-ghs_pop_gpw4_globe_r2015a [Dataset].

Kassié, D., Roudot, A., Dessay, N., Piermay, J.-L., Salem, G., and Fournet, F. (2017). "Development of a Spatial Sampling Protocol using GIS to Measure Health Disparities in Bobo-Dioulasso, Burkina Faso, A Medium-Sized African City." *International Journal of Health Geographics*, 16(1), 14. https://doi.org/10.1186/s12942-017-0087-7.

Kuffer, M., Owusu, M., Oliveira, L., Sliuzas, R., and van Rijn, F. (2022). "The Missing Millions in Maps: Exploring Causes of Uncertainties in Global Gridded Population Datasets." *ISPRS International Journal of Geo-Information*, 11(7), 403. https://doi.org/10.3390/ijgi11070403.

Lebakula, V., Epting, J., Moehl, J., Stipek, C., Adams, D., Reith, A., Kaufman, J., Gonzales, J., Reynolds, B., Basford, S., Martin, A., Buck, W., Faxon, A., Cunningham, A., Roy, A., Barbose, Z., Massaro, J., Walters, S., Woody, C., … Urban, M. (2024). "LandScan Silver Edition" [Data set]. Oak Ridge National Laboratory. https://doi.org/10.48690/1531770.

Meta and CIESIN - Columbia University. (2025). "High Resolution Settlement Layer (HRSL)." Available from: https://www.ciesin.columbia.edu/data/hrsl/. (Accessed: 08/04/2025).

Minaei, M. (2020). "Evolution, Density and Completeness of OpenStreetMap Road Networks in Developing Countries: The Case of Iran." *Applied Geography*, 119, 102246. https://doi.org/10.1016/j.apgeog.2020.102246.

National Statistical Service [Armenia], Ministry of Health [Armenia], and ICF. (2017). "Armenia Demographic and Health Survey 2015-16." Rockville, Maryland, USA: National Statistical Service, Ministry of Health, and ICF. https://dhsprogram.com/pubs/pdf/FR325/FR325.pdf.

Nieves, Jeremiah J., Maksym Bondarenko, Alessandro Sorichetta, Jessica E. Steele, David Kerr, Alessandra Carioli, Forrest R. Stevens, Andrea E. Gaughan, and Andrew J. Tatem. (2020). "Predicting Near-Future Built-Settlement Expansion using Relative Changes in Small Area Populations." *Remote Sensing*, 12(10): 1545. https://www.mdpi.com/2072-4292/12/10/1545.

Pettersson, Hans. (2023). "Sample Design and Sample Size for the Armenia MICS 2024." UNICEF Consultancy Report. Unpublished manuscript.

Qader, S. H., Lefebvre, V., Tatem, A. J., Pape, U., Jochem, W., Himelein, K., Ninneman, A., Wolburg, P., Nunez-Chaim, G., Bengtsson, L., and Bird, T. (2020). "Using Gridded Population and Quadtree Sampling Units to Support Survey Sample Design in Low-Income Settings." *International Journal of Health Geographics*, 19(1), 10. https://doi.org/10.1186/s12942-020-00205-5.

Qader, Sarchil H., Andrew Harfoot, Mathias Kuepie, Edith Darin, Sabrina Juran, Attila Lazar, and Andrew Tatem. (2022). "National Automatic preEnumeration Areas (preEAs) in Burkina Faso (2019), Version 1.0." WorldPop, University of Southampton. https://data.worldpop.org/repo/wopr/BFA/preEAs/v1.0/.

Qader, Sarchil H., Heather Chamberlain, Mathias Kuepie, Freja K. Hunt, Attila Lazar, and Andrew Tatem. (2023). "Field Testing of Pre-Enumeration Areas Created using Semi-Automated Delineation Approach, Democratic Republic of Congo." WorldPop, University of Southampton. https://dx.doi.org/10.5258/SOTON/WP00759.

Qader, Sarchil H., Veronique Lefebvre, Andrew J. Tatem, Utz Pape, Warren Jochem, Kristen Himelein, Amy Ninneman, Philip Wolburg, Gonzalo Nunez-Chaim, Linus Bengtsson, et al. (2020). "Using Gridded Population and Quadtree Sampling Units to Support Survey Sample Design in Low-Income Settings." *International Journal of Health Geographics*, 19: 1–16. https://dx.doi.org/10.1186/s12942-020-00205-5.

Qader, Sarchil, Veronique Lefebvre, Andrew Tatem, Utz Pape, Kristen Himelein, Amy Ninneman, Linus Bengtsson, and Tomas Bird. (2021). "Semi-Automatic Mapping of Pre-Census Enumeration Areas and Population Sampling Frames." *Humanities and Social Sciences Communications*, 8(1): 1–14. https://dx.doi.org/10.1057/s41599-020-00670-0.

Schiavina, M., M. Melchiorri, M. Pesaresi, P. Politis, S.M. Carneiro Freire, L. Maffenini, P. Florio, D. Ehrlich, K. Goch, A. Carioli, J. Uhl, P. Tommasi, and T. Kemper. (2023). "GHSL Data Package 2023." Publications Office of the European Union. https://dx.doi.org/10.2760/098587.

Stevens, Forrest R., Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem. (2015). "Disaggregating Census Data for Population Mapping using Random Forests with Remotely- Sensed and Ancillary Data." *PLoS ONE*, 10(2): e0107042. https://doi.org/10.1371/journal.pone.0107042.

Strano, Emanuele, Andrea Giometto, Saray Shai, Enrico Bertuzzo, Peter J. Mucha, and Andrea Rinaldo. (2017). "The Scaling Structure of the Global Road Network." *Royal Society Open Science*, 4(10): 170590. https://doi.org/10.1098/rsos.170590.

Thomson, D. R., Rhoda, D. A., Tatem, A. J., and Castro, M. C. (2020). "Gridded Population Survey Sampling: A Systematic Scoping Review of the Field and Strategic Research Agenda." *International Journal of Health Geographics*, 19(1), 34. https://doi.org/10.1186/s12942-020-00230-4.

Thomson, Dana R., Forrest R. Stevens, Nick W. Ruktanonchai, Andrew J. Tatem, and Marcia C. Castro. (2017). "GridSample: An R Package to Generate Household Survey Primary Sampling Units (PSUs) from Gridded Population Data." *International Journal of Health Geographics*, 16: 1–19. https://dx.doi.org/10.1186/s12942-017-0098-4.

United Nations (UN). (2005). "Designing Household Survey Samples: Practical Guidelines." In *Studies in Methods Series F No. 98*. Available from: https://unstats.un.org/unsd/demographic/sources/surveys/Handbook23June05.pdf. (Accessed: 08/04/2025).

United Nations Children's Fund (UNICEF). (2012). "Designing and Selecting the Sample." In *Multiple Indicator Cluster Surveys Round 4*. http://mics.unicef.org/tools?round=mics4. (Accessed: 08/04/2025).

United Nations Secretariat. (2007). "Report of the Sub-regional Workshop on Census Cartography and Management." ESA/STAT/AC.144/L.3.

Weber, E., Weaver, J., McKee, J., Lunga, D., Laverdiere, M., Swan, B., Yang, H., Urban, M., Cheriyadat, A., and Patlolla, D. (2017). "LandScan HD Afghanistan v1.0 Version v1.0)" [raster digital data]. Oak Ridge National Laboratory. https://doi.org/https://doi.org/10.48690/1524218.

WorldPop. (2025). "WorldPop Data." https://www.worldpop.org/datacatalog/. (Accessed: 08/04/2025).

Appendix A: Additional Figures

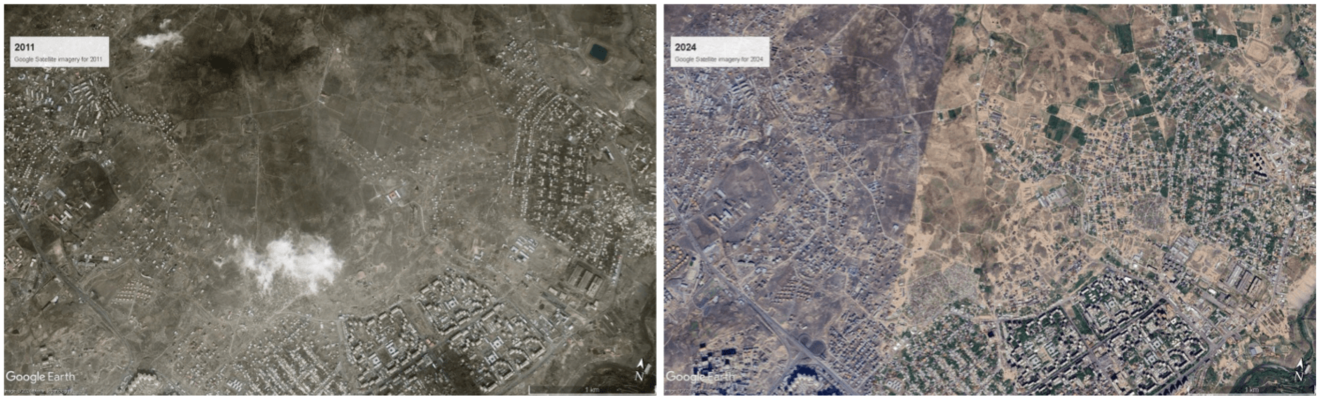
2011
Google Satellite imagery for 2011
Google Earth
1 km
N
2024
Google Satellite imagery for 2024
Google Earth
1 km
N

Figure A.1: Change in population distribution in an area between 2011 and 2024

Appendix B: Field and Offline Maps

To take advantage of the boundaries and other spatial information, the survey conductor can create field maps of selected pre-EAs for enumerators to feed their navigation when they collect the survey. Even in the presence of the field maps, it could be still challenging for enumerators to navigate themselves in the selected pre-EA especially, when the enumerators are not familiar with the area and cannot eyeball the boundaries from the information provided in the physical maps, like street address and some information about church and schools. A potential solution to this problem could be offline maps, which inform the enumerators of their location live and signal if they overstep outside the selected pre-EAs. In many developing countries like Armenia, access to the internet is a substantial challenge, especially in rural and remote areas, so enumerators can benefit from the offline maps that operate well with the minimum requirement of only access to satellite. So, enumerators should be able to navigate smoothly and avoid the risk of going beyond the boundaries of selected PSUs unless they are, for example, under the tunnel or in between narrow alleys in the mountain. In total, 400 field paper maps and 400 georeferenced offline maps were created using QGIS software.

Multiple settled areas are probably present in various pre-EAs. Map definitions of these settled areas may be helpful for ground navigation. Several actions have been taken to display the settled region on the field maps. Administrative boundaries with incorporated urban and rural areas were intersected with settlement boundaries. Zonal Statistic Polygons in QGIS were then used to determine the population sum for the resultant intersected polygons based on the gridded population data. A point layer was created from the output. Additionally, the settlement locations' maximum population values inside each pre-EA were extracted. The settlement with the highest population numbers is shown by the black circle surrounding settlement sites on the field map. This can help the enumerator to find the most densely populated area as a starting point. The X and Y coordinates were calculated for all the settlement points and were shown on the field map (Figures B.1 and B.2).

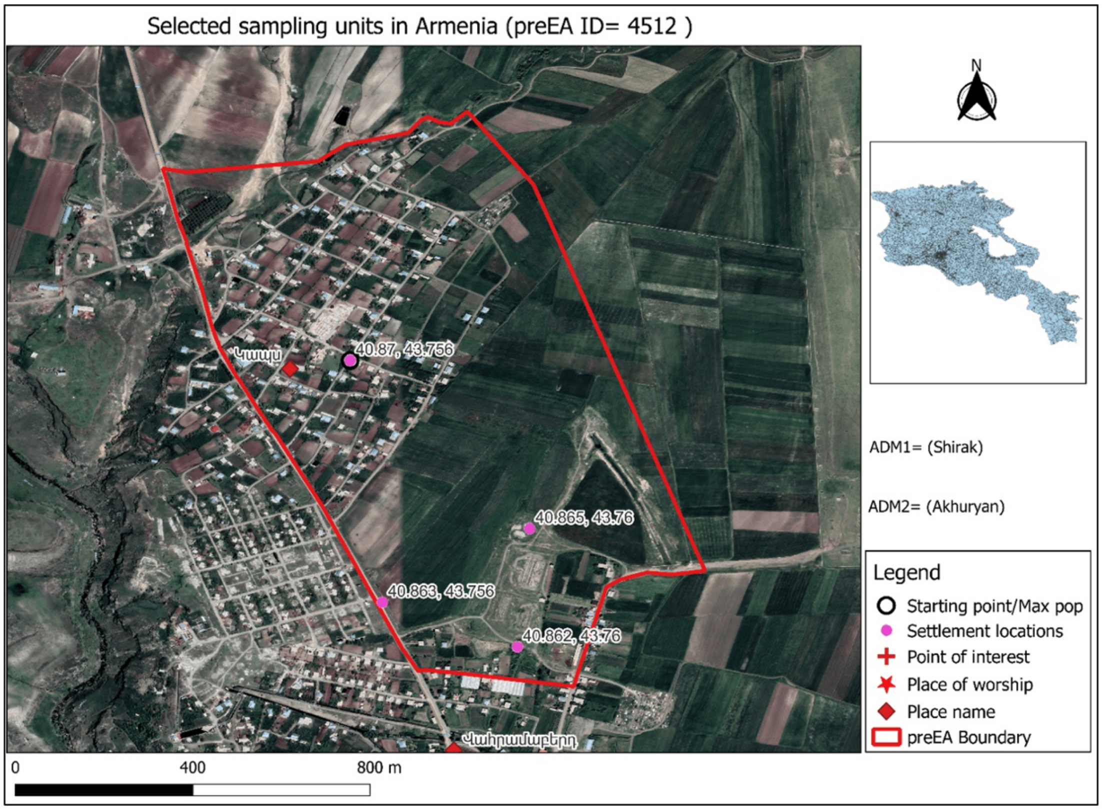

Figure B.1: An example of a detailed field paper map in rural areas

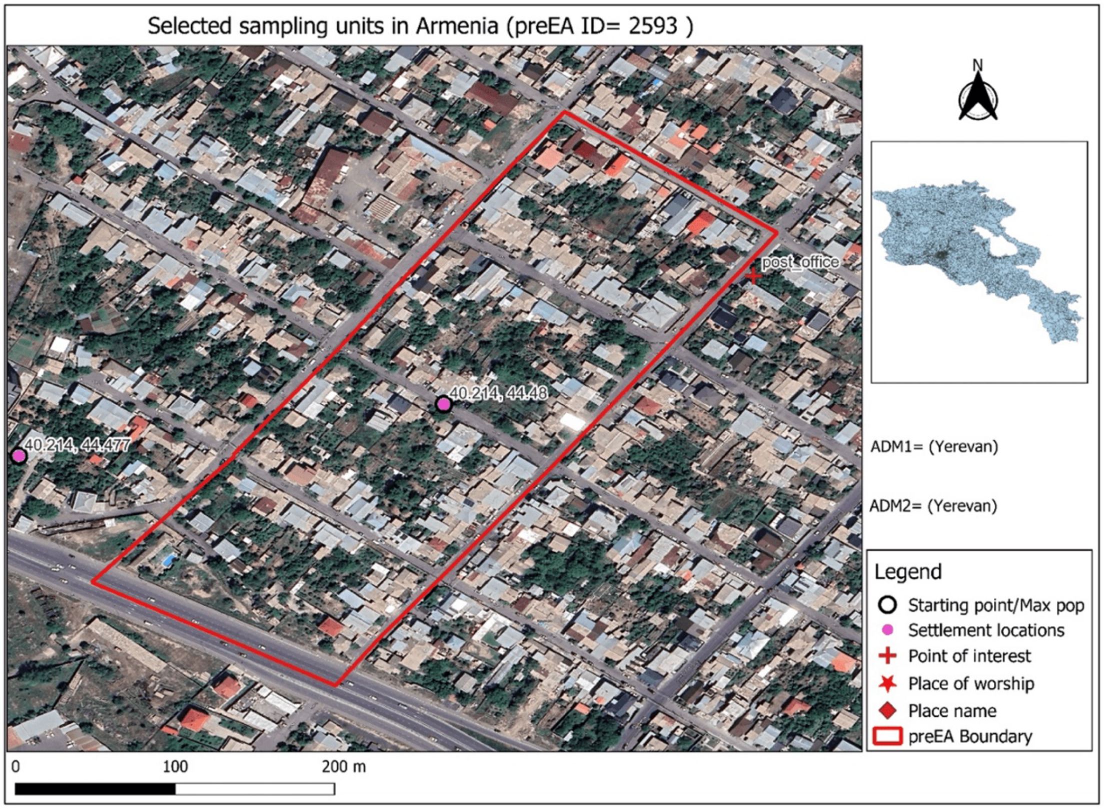

Figure B.2: An example of a detailed field paper map in urban areas

Appendix C: Survey Weights

Sampling weights account for the fact that different members of the population have different probabilities of being selected for interviews, represent various numbers of people in the overall population, and are necessary when computing the representative statistics at the level of the domain. If required, sampling weights are also adjusted to account for non-responsive rates given the survey design.

The dataset will have two sets of weights, including household and individual weights. The household weights are the inverse probability of selection of households and are calculated from the following two components in our two-stage sampling design. The first component is the sampling weight (inverse probability of selection) of PSU within the stratum, and the second component is the sampling weight of selection of households within the PSU. For calculating individual weights, the third component, sampling weights of individuals within the household, is added to compute the household weights. Each of the components is calculated as follows:

Component 1: The inverse probability of selection of PSU within the stratum by using PPS (Probability Proportional to Size) is calculated as:

$$W_{psu} = \frac{1}{P_{psu}} = \frac{N_{stratum}}{n_{psu} \times N_{psu}},$$

where W_{psu} is the sampling weight of PSU within the stratum, P_{psu} is the probability of selection of PSU within the stratum, n_{psu} is the number of selected PSUs within the stratum, N_{psu} is the size of selected PSU, and $N_{stratum}$ is the population of the stratum. The size measure can be the population, the number of households, the number of electors, or the school attendance, while the number of households would be the preferred option in most household surveys.

Component 2: The inverse probability of selection of household within PSU is calculated as:

$$W_{hh_{psu}} = \frac{1}{P_{hh_{psu}}} = \frac{N_{hh_{psu}}}{n_{hh_{psu}}},$$

where $W_{hh_{psu}}$ is the sampling weight of household within PSU, $P_{hh_{psu}}$ is the probability of selection of household within PSU, $n_{hh_{psu}}$ is the number of sampled (interviewed) households within PSU, and $N_{hh_{psu}}$ is the total number of households within PSU.

Component 3: The inverse probability of selection of individual within the household is calculated by:

$$W_{ind_{hh}} = \frac{1}{P_{ind_{hh}}} = \frac{N_{ind_{hh}}}{n_{ind_{hh}}} = \frac{N_{ind_{hh}}}{1},$$

where $W_{ind_{hh}}$ is the sampling weight of individual within the household, $P_{ind_{hh}}$ is the probability of selection of individual within the household, $n_{ind_{hh}}$ is the number of sampled (interviewed) individuals within the household, equal to 1, as only one individual was allowed to be interviewed from each household, and $N_{ind_{hh}}$ is the size of the household surveyed (asked and recorded during the interview).

Based on these components, household and individual weights are calculated as:

$$W_{hh} = W_{psu} \times W_{hh_{psu}},$$

$$W_{ind} = W_{hh} \times W_{ind_{hh}}.$$