December 2025

Technical Guide: Adaptive Cluster Sampling for Urban Micro and Small Enterprises

AUTHORS | Edoardo Totolo
Paul Gubbins

CENTER for FINANCIAL INCLUSION | ACCION

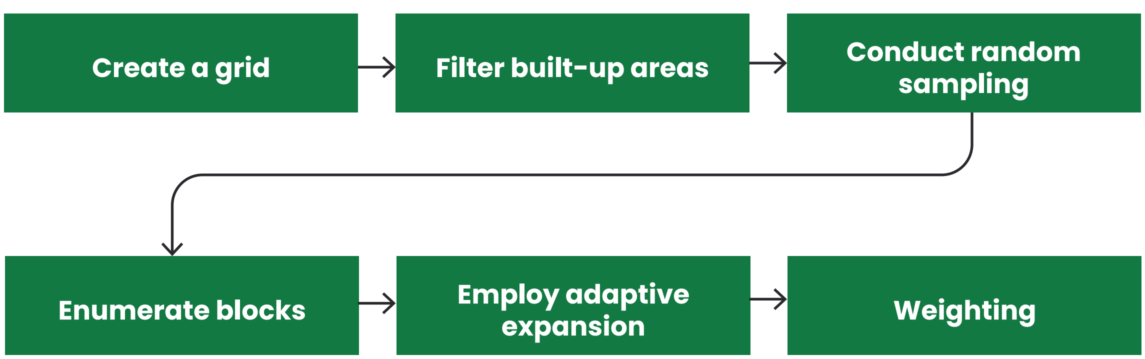
01 Introduction

This technical guide documents the sampling methodology used in the *Small Firms, Big Impact* study, conducted by the Center for Financial Inclusion (CFI) with support from Mastercard Center for Inclusive Growth. The study aims to produce representative and comparable data on micro and small enterprises (MSEs) in five large metropolitan areas: São Paulo (Brazil), Addis Ababa (Ethiopia), New Delhi (India), Jakarta (Indonesia), and Lagos (Nigeria).

Across these cities, MSEs form the backbone of employment and income generation but remain largely invisible in official data. Many operate informally or from home and are absent from tax records or business registries. The methodology was developed to fill this gap by using geospatial block enumeration with adaptive cluster sampling (ACS) to systematically capture the full diversity of urban small enterprises, both formal and informal.

The guide is intended for research teams, statistical agencies, and development practitioners who wish to replicate or adapt the approach for other contexts and research objectives. It describes each stage of the sampling design, explains the rationale behind key methodological decisions, and provides guidance for implementation, weighting, and data quality management.

FIGURE 1. OVERVIEW OF THE SAMPLING PROCESS:



02 Conceptual Framework

2.1. MOTIVATION, RATIONALE AND DESIGN PARAMETERS

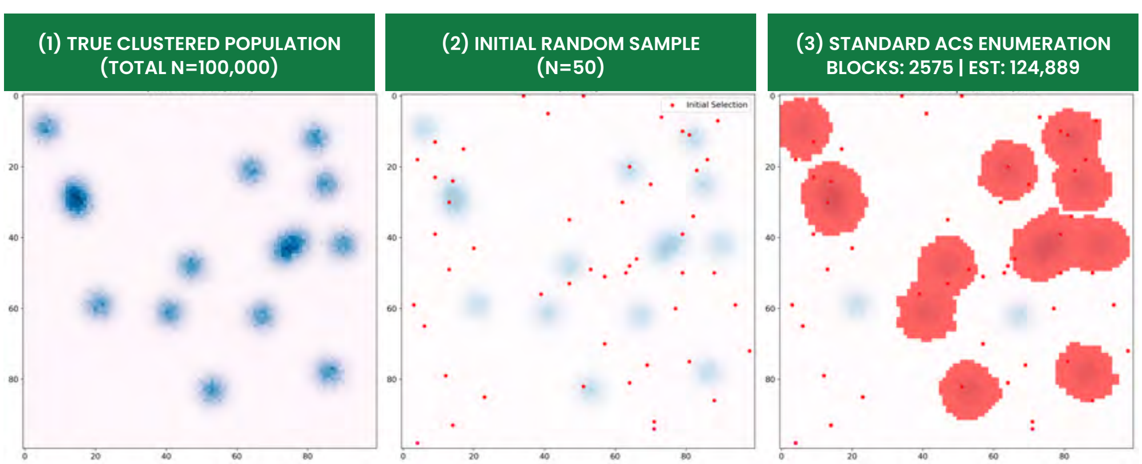
Traditional enterprise surveys rely on pre-existing lists or registries of firms. In low- and middle-income economies, such lists are incomplete, excluding large numbers of informal businesses that operate outside formal registration. As a result, survey samples are often biased toward visible, better-established firms, while the vast majority of micro and household enterprises remain underrepresented.

The approach addresses this limitation by grounding the sample in geography rather than institutions. Instead of sampling from a list of businesses, the design samples small geographic units of the urban fabric, referred to as blocks, and directly enumerates the enterprises located within them to create an 'on-the-fly' sampling frame of businesses from which a random sample is selected for the survey.

Adaptive cluster sampling then allows the survey fieldwork to expand dynamically wherever business activity is concentrated. Because small businesses can be clustered geographically, a simple random sampling design can be inefficient as it expends significant fieldwork effort enumerating empty blocks. ACS exploits this clustering by using the discovery of a business to trigger expanding the sample into adjacent areas, thereby concentrating resources on high-density "hotspots."

FIGURE 2. SIMULATION OF THE BLOCK ENUMERATION AND ACS METHODOLOGY

A synthetic population of 100,000 businesses clustered in a grid of 100 by 100 blocks (10,000), an initial sample of 50 blocks with an expansion threshold of 10 businesses is used to estimate the total number of businesses in the population.



Block enumeration with ACS can be a robust alternative compared to other methodologies for the following reasons:

- **Unbiased Population Estimates:** Unlike methods where selection probabilities can be opaque, ACS allows for the calculation of inclusion probabilities for every sampled unit, albeit these inclusion probabilities are more complex to calculate than for a simple random sample of primary geographic sampling units. This enables estimation of unbiased population totals and means, making the survey results statistically representative of the defined geographic area.
- **Efficiency in Capturing "Rare" and "Clustered" Populations:** Small or informal businesses are often "rare" (sparse in many areas) but "clustered" (dense in specific hotspots). ACS exploits this by triggering intense sampling only where businesses are actually found. Aga et al. (2023) demonstrate that for a fixed level of desired precision (standard deviation), ACS requires significantly less fieldwork effort (fewer enumerated blocks) than a comparable simple random sample.
- **Defined Population of Inference:** ACS forces the researcher to clearly delineate the universe of interest as a specific geographic area (e.g., a set of urban blocks). This contrasts with methods that may lack a rigorous sampling frame, ensuring that the resulting data can be extrapolated to a clear, finite population.
- **Quality Control via Technology:** The methodology encourages the use of GPS and path-tracking software to ensure enumerators fully cover the selected block areas (BAs), reducing the risk of enumerator discretion or path deviation that can bias other methods such as random walks.

However, block enumeration with ACS does have a number of practical challenges from an implementation standpoint. It involves several design parameters, and the fact that the total number of blocks that will be enumerated cannot be known in advance requires managing these parameters carefully, especially in circumstances with a fixed fieldwork budget (Table 1).

TABLE 1

Design Choice	Description	Considerations & Recommendations
1. **Geographic Boundaries (N)**	Defining the total universe of the study area (the sampling frame)	↗ **Sources:** Use administrative boundaries, natural boundaries, or supplementary data like night lights. ↗ **Recommendation:** Official administrative boundaries can have benefits such as aligning with existing datasets and political jurisdictions but can be outdated in terms of covering areas of actual population density; Consider using remote sensing data to complement boundary setting.

Design Choice	Description	Considerations & Recommendations
2. **Block Area (BA) Size**	The physical dimensions of the block - the Primary Sampling Unit (PSU)	↗ **Recommendation:** A size of 150m x 150m is the standard used in World Bank Informal Enterprise Survey program and was adopted by CFI. ↗ **Trade-off:** Must be small enough for a single team to enumerate in a reasonable time, but large enough to meaningfully capture clusters of businesses.
3. **Initial sample of blocks (n)**	The number of blocks (PSUs) to randomly select in the first stage	↗ **Impact:** Must be small enough for a single team to enumerate in a reasonable time, but large enough to meaningfully capture clusters of businesses.
4. **Expansion Threshold (C)**	The number of target businesses discovered in a block that triggers the enumeration of neighbors	↗ **Impact:** Higher thresholds reduce fieldwork effort (fewer expansions) but may miss smaller clusters. ↗ **Recommendation:** Start with a conservative (higher) value for C to control budget/effort. You can calibrate this after enumerating a few blocks.
5. **Neighborhood Definition**	The rule defining which "adjacent" blocks are added to the sample when C is met	↗ **Standard Practice:** Second-order: The original block plus all 8 surrounding squares in the first expansion, and all adjacent blocks (except for already enumerated blocks) in subsequent expansions. ↗ **Alternatives:** First-order (4 adjacent squares) or strip sampling (linear) can also be used depending on the expected spatial pattern of businesses.
6. **Stratification**	Dividing the geographic area into distinct categories (strata) before sampling	↗ **Impact:** Improves efficiency and reduces variance ↗ **Data Sources:** Can use Nighttime Lights satellite imagery or land-use maps to stratify into high/low density areas. ↗ **Categories:** Useful categories include Residential, Commercial/Industrial, Market Centers, and Mixed use.
7. **Target Population (j)**	The operational definition of an "eligible" business	↗ **Definition:** CFI defined eligible businesses primarily with a size criteria: only enumerated firms with 0-9 regular employees were eligible; this will depend on the specific objectives of the study. ↗ **Context:** Definitions based on other criteria (such as formal vs. informal) should be anchored to local laws (e.g., specific municipal licenses).
8. **Second-Stage Selection & probability (p)**	The method and probability (p) for selecting enumerated businesses for a detailed interview	↗ **Method:** Eligible businesses are randomly selected from the enumerated eligible list in real-time using CAPI software. ↗ **Weighting:** The inverse of this probability (1/p) drives the second-stage weight.

Design Choice	Description	Considerations & Recommendations
9. **Controlling Fieldwork Effort**	Mechanisms to prevent the sample from exploding due to massive clusters	↗ **Risk:** Since the final sample size is unknown a priori in ACS, costs can spiral. ↗ **Strategy:** Can be managed by adjusting the Expansion Threshold (C) by conducting a calibration phase with the initial sample. For a fixed number of target surveys/interviews, raising the selection probability of p will also limit total fieldwork.
10. **Quality Control**	Monitoring enumerators to ensure true full enumeration of blocks	↗ **Technology:** Use GPS path tracking software to ensure enumerators physically walked all paths within the 150x150m boundary. Can also be monitored by capturing the GPS coordinates of enumerated businesses.

The sections that follow describe CFI's implementation of the block enumeration with ACS methodology for its *Small firms, Big Impact study*, reflecting design choices based on the time and available resources for the study. Other implementers may make different design choices based on their own research context and budget. Where applicable, we explain our decision choices and lay out alternatives others could consider.

03 Constructing the Geospatial Sampling Frame

3.1 DEFINING THE SAMPLING BOUNDARY

The first step in constructing the sampling frame is to delineate the boundary of the sampling area. To align with established political boundaries and facilitate use of additional data sources (such as census-derived population data), official administrative boundaries were used to define the sampling boundaries in the study: GADM v4 for São Paulo, Addis Ababa, and Jakarta, and UN HDX for Delhi and Lagos (Table 2).

TABLE 2. DEFINING THE SAMPLING BOUNDARY

City	Level-2 Administrative units used for sampling boundary	Total area covered by sampling boundary (km2)	Total number of 'built' blocks (150m x 150m) in final sampling grid	Total area of 'built' blocks in final sampling grid (km^2)	Details
Addis Ababa, Ethiopia	Addis Ababa chartered city	553	15,842	361	
Jakarta, Indonesia	Five of six administrative cities in Jakarta Province*: Jakarta Barat, Jakarta Pusat, Jakarta Selatan, Jakarta Timur, Jakarta Utara	650	26,293	593	*The Kepulauan Seribu regency which is comprised of small islands off the mainland is excluded.
Delhi, India	All eleven districts in Delhi state: North, North West, North East, Central, Shahdara, East, South East, South, New Delhi, South West, West	1,925	30,880	893	

City	Level-2 Administrative units used for sampling boundary	Total area covered by sampling boundary (km2)	Total number of 'built' blocks (150m x 150m) in final sampling grid	Total area of 'built' blocks in final sampling grid (km²)	Details
São Paulo, Brazil	São Paulo Municipality	1,814	38,017	1,015	
Lagos, Nigeria	Five of twenty local government areas in Lagos State*: Lagos Mainland, Lagos Island, Eti-Osa, Apapa, Surulere	266	9,282	208	* These five LGAs correspond to those in the Lagos [Eko] administrative division of Lagos state: https://lagosstate.gov.ng/about-lagos/

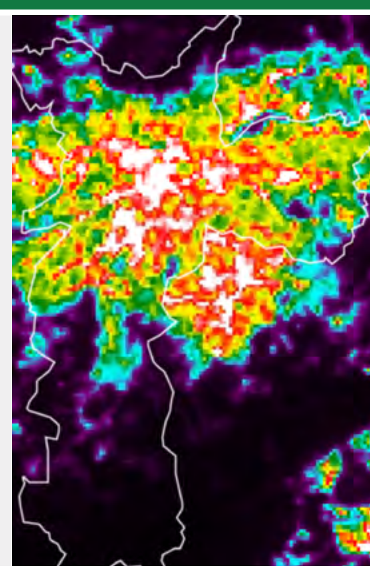
However, city administrative boundaries can omit areas of peri-urban or industrial zones where many small firms operate. In São Paulo, for example, night-light data revealed significant built-up extensions beyond the official municipal boundary. High resolution satellite imagery (such as Sentinel-2) or NASA VIIRS Night Light can be used as an alternative way to delineate the natural extent of economic activity at the city level.

FIGURE 3. ADMINISTRATIVE BOUNDARIES AND BUILT-UP FOOTPRINT OF SÃO PAULO BASED ON VISUAL SPECTRUM AND NIGHT-LIGHT INTENSITY.

TRUE-COLOR DAY-TIME SATELLITE COMPOSITE

Source: Cloud-free mosaic of Sentinel-2 imagery via Sentinel-Hub

NIGHTTIME LIGHTS (NTL)



Source: NASA VIIRS (lack Marble) via Google Earth Engine

3.2. CREATING THE SAMPLING GRID

Once the sampling boundary is defined, the area is divided into a regular grid of 150 m × 150 m blocks, equivalent to 2.25 hectares each. These blocks form the primary sampling units (PSUs). A custom python script generates the grid, assigning a unique identification code and centroid coordinate to every block.

FIGURE 4. EXAMPLE OF INITIAL GRID OF 69,173 BLOCKS COVERING THE SÃO PAULO STUDY BOUNDARY

3.3. FILTERING BUILT-UP BLOCKS

To avoid sampling empty or uninhabited areas such as bodies of water, farmland, or forest, only blocks that contain built structures are retained. Built-up areas are identified using the Dynamic World v1 land-cover dataset (10 m resolution) accessed via Google Earth Engine.

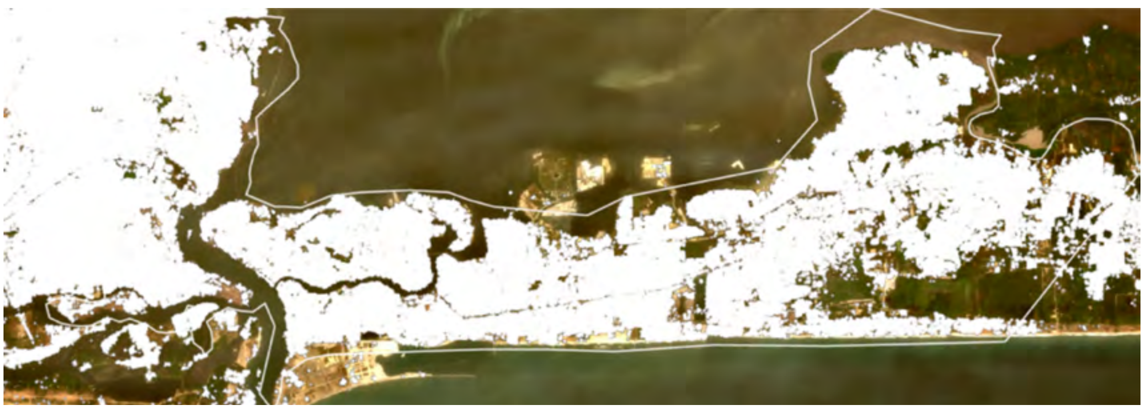
A cloud-free composite of the most recent four months of imagery is produced, and for each pixel the modal land-cover classification is extracted. The share of "built" pixels within each 150 m block is then computed, and blocks with at least 75 percent built-pixel composition are included in the final sampling frame. Since this step requires processing pixels over a large area using Google Earth Engine, large metropolitan areas are processed in segments of roughly 2,000 blocks per batch to remain within the processing limits.

In Lagos, additional refinement was necessary because dense informal settlements were misclassified as "bare" in the Dynamic World land-cover dataset. To correct this, the Open Buildings v3 dataset was cross-referenced, and any block containing at least one structure with a confidence score above 0.75 was retained, resulting in a more accurate built-up mask.

FIGURE 5. BUILT-UP AREA CLASSIFICATION FOR LAGOS USING DYNAMIC WORLD (RED = BUILT) AND OPEN BUILDINGS OVERLAYS

A: Land-use/land-cover (red = built pixels), Source: Dynamic World

B: Building boundaries (white = building outlines), Source: Open buildings



C: Close up of area with poor correspondence of built-up classification and presence of buildings

The result of this step is a spatially explicit grid of built blocks that collectively define the eligible survey universe.

FIGURE 6. FINAL SAMPLING GRID COMPRISED OF 92,282 BLOCKS WITH AT LEAST ONE DETECTED "HIGH CONFIDENCE" BUILDING

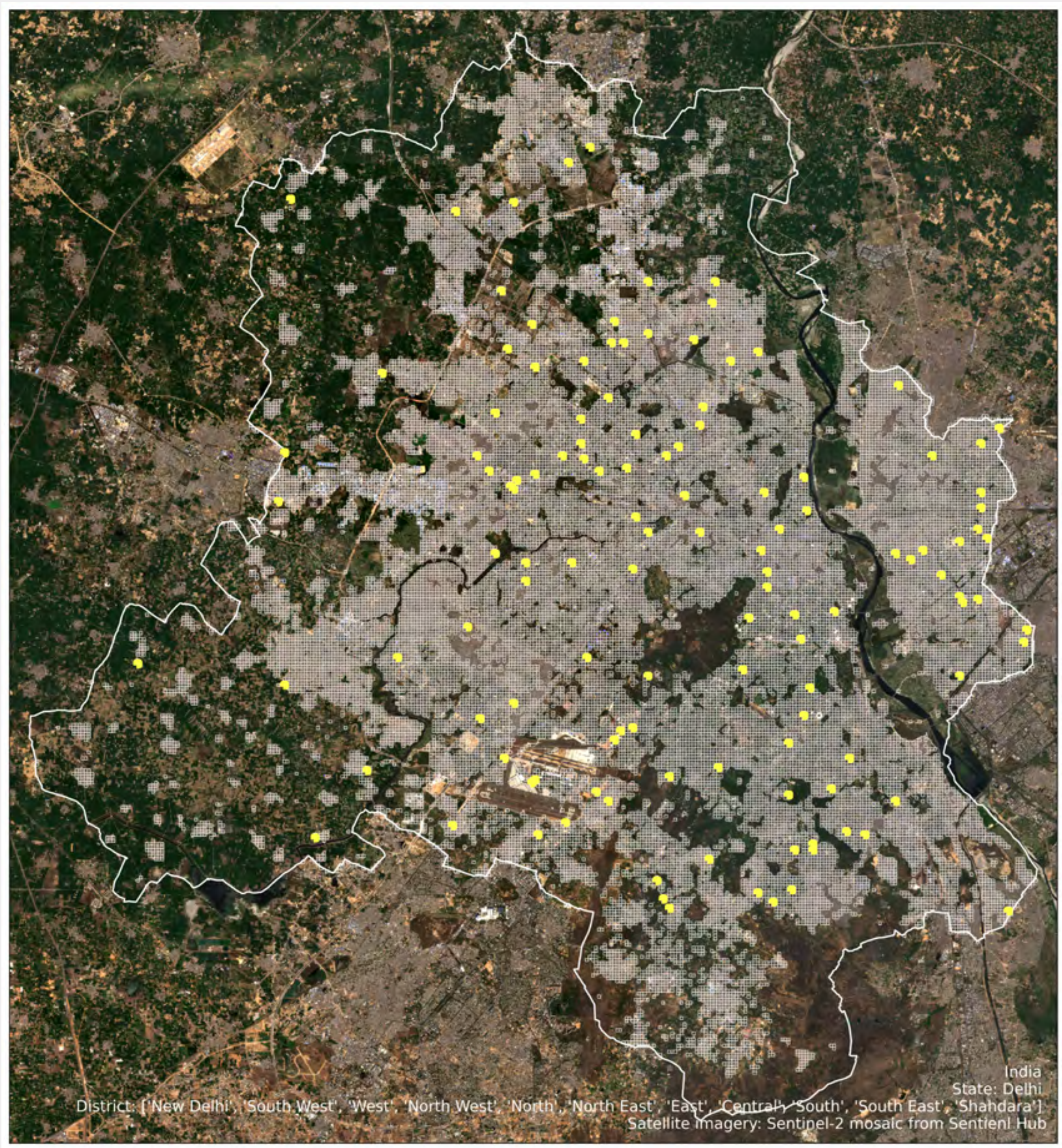
3.4. SELECTING THE INITIAL SAMPLE OF BLOCKS

From the final built-up grid, an initial simple random sample of 125 blocks was drawn for each city. The number of initial blocks can be adjusted according to budget and desired precision but should be sufficient to ensure geographic coverage and to allow adaptive expansion. The main trade-off in this decision is that for a given underlying geographic distribution of businesses and expansion threshold, the larger the initial sample size, the greater the expected fieldwork but with the benefit of increasing statistical precision.

For this study, we allowed the firm implementing the survey the opportunity to review the 125 initial sampled blocks to screen urban blocks that were off-limits to the public for several reasons (including security concerns in São Paulo). Of the remaining blocks, we randomly selected 100 blocks for the final sample. The sampling grid with a tag identifying the initial sample of blocks can be exported as a polygon shapefile to be loaded into field tablets for navigation and enumeration.

One variation to consider at this stage is stratification, which is used by the World Bank Informal Enterprise Survey program. The sampling grid can be divided into different categories (such as Residential, Commercial/Industrial, Market Centers, Open Areas/Inaccessible) and selecting blocks randomly within each stratum. This approach improves statistical precision because stratification ensures that areas with high density are adequately represented in the initial draw, minimizing the risk of missing major clusters.

FIGURE 7. RANDOMLY SELECTED 125 BLOCKS FOR ENUMERATION IN NEW DELHI



04 Field Enumeration and Adaptive Expansion

4.1. BLOCK ENUMERATION

Enumerators visit each sampled block in the field and record all visible businesses operating within the 150 m × 150 m area based on the following eligibility criteria:

- Households (with evidence of having a business operating from them, e.g. from signage or clear activity visible from the road)
- Non-households with a permanent structure
- Non-households with a semi-permanent structure, including stalls and stands

Business NOT eligible for enumeration include the following:

- Street vendors and hawkers operating at a temporary location without any fixed premise
- International branded stores in high-end malls operating out of permanent or semi-permanent structures (temporary mobile vendors, for example, are excluded)

For each identified business, enumerators record:

- main activity or sector;
- number of employees; and
- structural characteristics (commercial-permanent, commercial-semi-permanent, household).

Only enterprises with 10 or fewer employees, including the owner, are eligible for the main survey. Enumerators use tablets equipped with GPS to mark business locations and ensure full spatial coverage of the block.

4.2. APPLYING ADAPTIVE CLUSTER SAMPLING

The study applied the ACS procedure initially developed by Thompson (1990) and applied by the World Bank in the informal enterprise surveys to identify contiguous zones of high business density.

- **Expansion threshold (E):** After an initial calibration phase, the following expansion thresholds were used: Nigeria (15), Delhi (25), Jakarta (25), Addis Ababa (5), São Paulo (5).
- **Expansion rule:** If a sampled block contains ≥ E eligible businesses, all eight neighboring blocks (sharing sides or corners) are enumerated.
- **Recursion:** If any neighboring block also meets E, the process continues outward.
- **Stopping rule:** Expansion ceases when no new adjacent block meets the threshold. For thestudy, we introduced an additional stopping rule: If 10% of the total target number of interviews have been completed within the network originating from the initial sampled block, the expansion process stops. The rationale was to have a failsafe to ensure that the interview data would not be

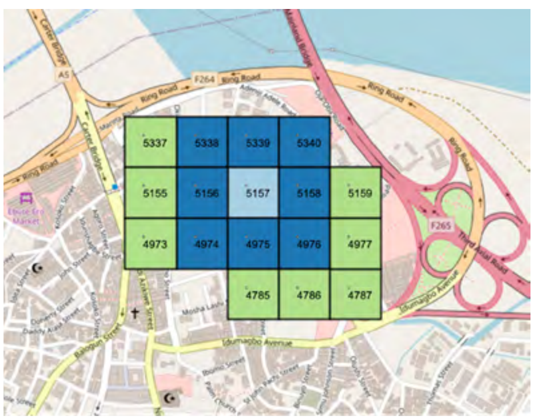
overly concentrated in a single or small number of clusters[1].

Each set of connected blocks meeting the threshold constitutes a network. Blocks that do not trigger expansion remain as single-block clusters. This process allows the sample to “grow” around real concentrations of enterprise activity such as markets, commercial corridors, or industrial estates, without the need for a full-city census.

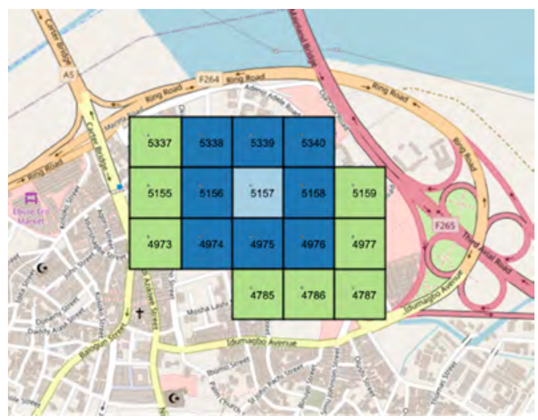
In Lagos, for example, an initial sampled block exceeded the threshold with 25 eligible businesses, triggering enumeration of its eight neighbors. Two of those also exceeded the threshold, expanding the network to 19 blocks before reaching saturation.

FIGURE 8. ILLUSTRATION OF ADAPTIVE EXPANSION SEQUENCE: BLUE = INITIAL BLOCK; DARK GREEN = FIRST EXPANSION; LIGHT GREEN = SECOND; PINK = FINAL EXPANSION.

A Block 5157 (light blue) is a block in the initial sample and 25 eligible businesses are discovered here. Since 25 exceeds the expansion threshold (E) of 15, all adjacent blocks (dark blue) are enumerated. After enumeration it is found that block 5156 and 4976 have 16 and 30 businesses, respectively, so all their adjacent blocks (light green) are enumerated.



B After enumerating the light green blocks it is found that the total number of eligible business in blocks 4787 and 5155 is 15 and 18, respectively, so all their adjacent blocks are enumerated (dark green).



4.3. SELECTING BUSINESSES FOR THE MAIN SURVEY

As eligible businesses are identified during enumeration, a random selection is made by the CAPI software to select eligible business for the main survey. We used a selection probability of 1 in 3.

1 A simulation to understand the impact of this stopping rule using a synthetic population of businesses suggests that sample estimates of the total population of businesses remain unbiased, but the standard error of the estimate increases substantially.

05 Weighting

The core of the ACS methodology relies on calculating an exact inclusion probability for every network of blocks found. The inclusion probability isn't about how dense the network is; it is strictly about how much space it takes up in the sampling grid (its footprint). A cluster with a larger footprint (more network blocks) is harder to miss, so it has a higher probability of inclusion and, consequently, a lower weight in the final analysis. This inclusion probability is calculated as a function of the total number of blocks in the sampling grid, the initial number of blocks sampled, and the size of the network found.

The CFI study used the following set of equations to calculate survey weights.[2] Let (j) denote an individual business within cluster (k) originating from an initial block (i). The sampling weight for business (j) is defined as:

EQUATION 1

$$Sampling\ weight_j = \frac{1}{\pi_{i,k} \times \delta_{j,k}}$$

where

$\pi_{i,k}$=Probability block i in cluster k is included in sample
$\delta_{j,k}$=Probability eligible business j in cluster k is interviewed

EQUATION 2

$$\pi_{i,k} = 1 - \frac{\binom{N-m_{i,k}}{n}}{\binom{N}{n}}$$

where

N=Total number of blocks in sampling frame

n=Initial number of sampled blocks

$m_{i,k}$=number of blocks in cluster k to which block i belongs

Notes:

$\pi_{i,k}$ *is* constant for all blocks in any given cluster k

In the case that a starting block *i* does not meet the threshold, it is considered as a cluster with a size of $m_{i,k}=1$

2 The standard Horvitz-Thompson estimator differs from the equation presented here in that the total number of edge units in the network are subtracted from the cluster size. The inclusion probability of a network is calculated based on the number of units within the network that could trigger its selection. The edge units do not affect this probability. Edge units are observed (measured) to confirm the boundaries of the network but their associated data value (which is zero or below the threshold) has a zero contribution to the standard estimators for the population total or mean.

EQUATION 3

$$\delta_{j,k} = \frac{I_k}{B_k}$$

where

I_k=Total number of businesses interviewed in cluster k
B_k=Total number of eligible businesses enumerated in cluster k

Weights are computed using the cluster information exported from the field enumeration tool and applied during statistical estimation to produce city-level population inferences.

06 Quality Control and Data Management

Quality assurance is central to maintaining methodological integrity. Each city team should follow a standardized set of procedures:

1. **Training:** Enumerators and supervisors should complete a training combining classroom instruction and field practice. Session should covered eligibility screening, block navigation, and the adaptive expansion rules.
2. **Supervision:** Daily monitoring of enumeration maps ensure all blocks are fully covered and prevent duplication at boundaries.
3. **Verification:** A percentage of enumerated blocks should be re-visited by supervisors for validation.
4. **Data management:** Enumeration data should be uploaded daily to a central server. Automated scripts check for outliers (for instance, blocks reporting implausibly high numbers of enterprises).
5. **Documentation:** Each country team should produce a Sampling Logbook containing shapefiles of blocks, expansion maps, and weighting files.

07 Summary

The methodology adopted in *Small Firms, Big Impact* represents an improvement over traditional MSME survey designs that rely on incomplete business lists or random walks. By anchoring sampling in spatial data and allowing expansion around real clusters of activity, it produces replicable, representative estimates of city-level population parameters among small businesses.

However, the method does require technical preparation for weighing trade-offs among several design choices, use of reliable satellite or geospatial data, and close field supervision. In addition, the total fieldwork effort and therefore the exact costs of implementing adaptive sampling cannot be known with certainty in advance, requiring flexibility and appropriate budgeting mechanisms. In addition, while this approach is well suited to the study of small businesses, it is not well suited for the study of larger businesses that may operate in off-limits industrial zones, or business parks, or small businesses that do not have a physical retail presence, such as online retailers.

Despite these challenges, the gains in representativeness and comparability make this approach an efficient and robust option for studying small enterprises in complex urban settings.

08 Technical Resources

Several code repositories that were generated for the different parts of this project are publicly available and can be used for adaptation for other studies and contexts:

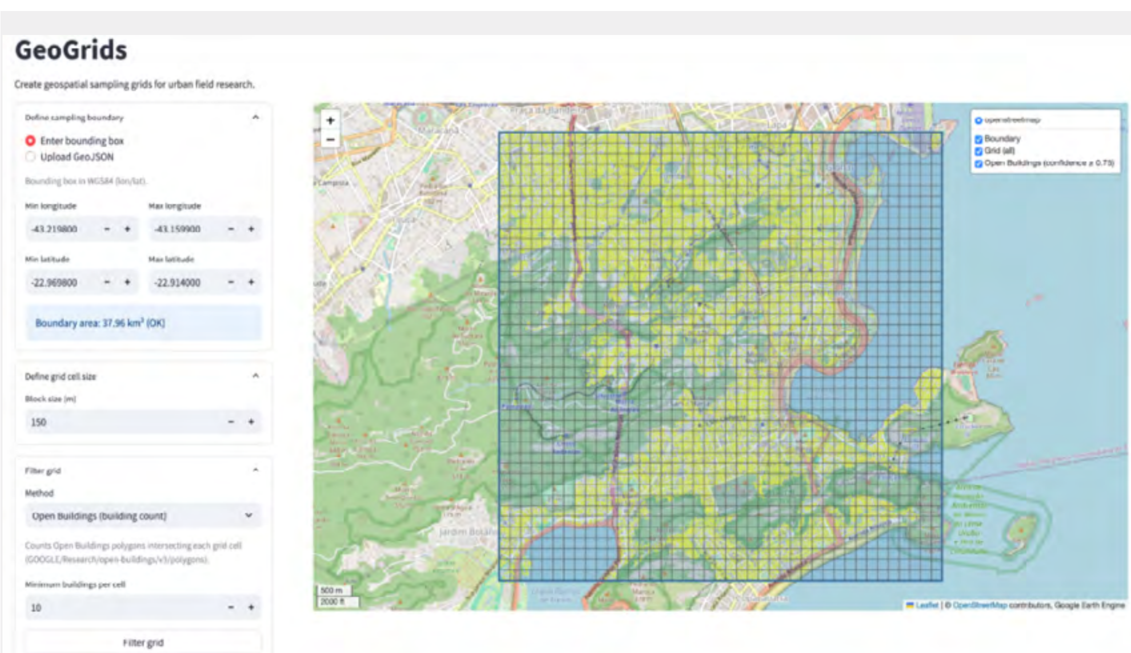
1. **Sampling grids:** https://github.com/Center-for-Financial-Inclusion/cfi-map2-sampling-grids

This repository contains all the data and code necessary to reproduce the "Stage 1" geospatial sampling grids for the five cities included the study, namely: São Paulo (Brazil), Addis Ababa (Ethiopia), Jakarta (Indonesia), Delhi (India) and Lagos (Nigeria). The methodology for each city is described in detail in the /documentation folder. This includes a notebook (https://github.com/Center-for-Financial-Inclusion/cfi-map2-sampling-grids/blob/main/notebooks/expand_block.ipynb) which provides a set of utilities to view sampled blocks and visualize potential expansions of a network if Adaptive Cluster Sampling expansion thresholds are met.

```
]:  ##########################################
    # Example 2: Visualize all the adjacent blocks and block ids for two expansion 'rounds'
    ##########################################

    origin_gdf = sampling_grid[sampling_grid['block_id'] == origin_block_id]
    adjacent_blocks_gdf = get_adjacent_blocks(origin_block_id, sampling_grid)
    network = pd.concat([origin_gdf, adjacent_blocks_gdf])

    edge_block_id = '4974'
    new_adjacenet_blocks_gdf = get_new_adjacent_blocks(edge_block_id, network, sampling_grid)
    new_network = pd.concat([network, new_adjacenet_blocks_gdf])

    # Plot OSM base
    plot_osm_base(new_network)
    # Plot the new adjacent blocks
    new_adjacenet_blocks_gdf.boundary.plot(ax=plt.gca(), color="red", linewidth = 2.75, facecolor = "red", alpha = 0.
    # Plot the adjacent blocks
    adjacent_blocks_gdf.boundary.plot(ax=plt.gca(), color="orange", linewidth = 2.75, facecolor = "orange", alpha = (
    # Plot the origin block
    origin_gdf.boundary.plot(ax=plt.gca(), color="green", linewidth = 4, aspect = 1, facecolor = "green", alpha = 0.5
    # Display block ids of blocks in the network
    for idx, row in new_network.iterrows():
        plt.text(row.geometry.centroid.x, row.geometry.centroid.y, row['block_id'], fontsize=11, color = "black", ha=

(6.468797000001055, 6.455454000003827, 3.396383999997767, 3.382961000002905)
```

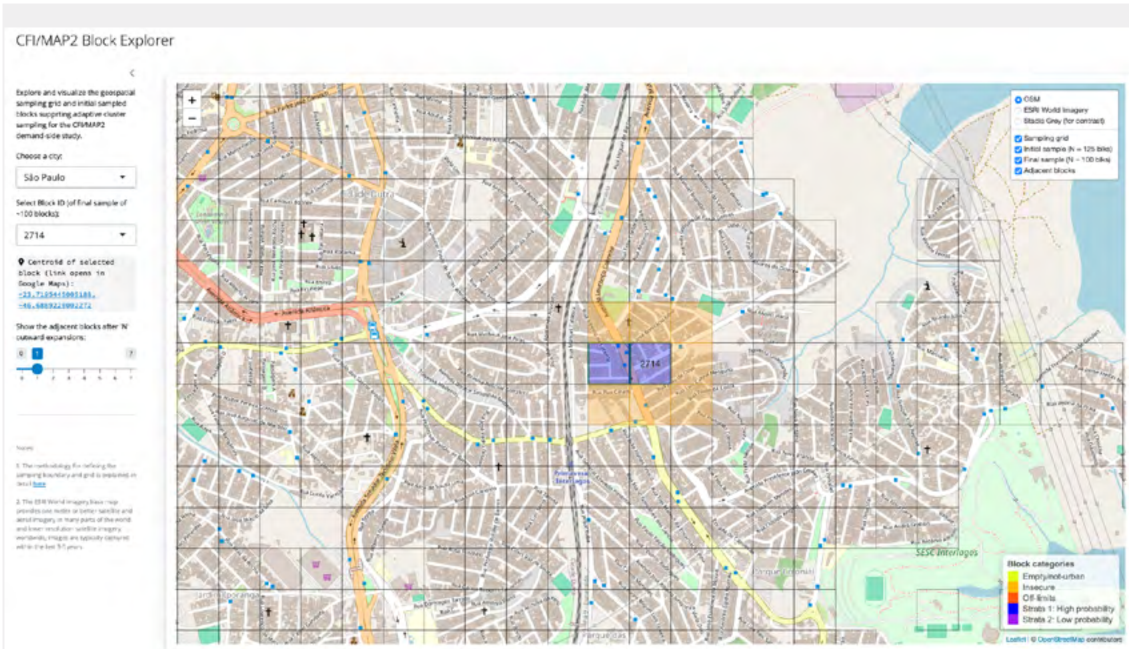
2. Geogrids app: https://github.com/Center-for-Financial-Inclusion/geogrids-app

This repository contains python code for a streamlit app that creates custom geospatial grids using the approach used in the CFI study.



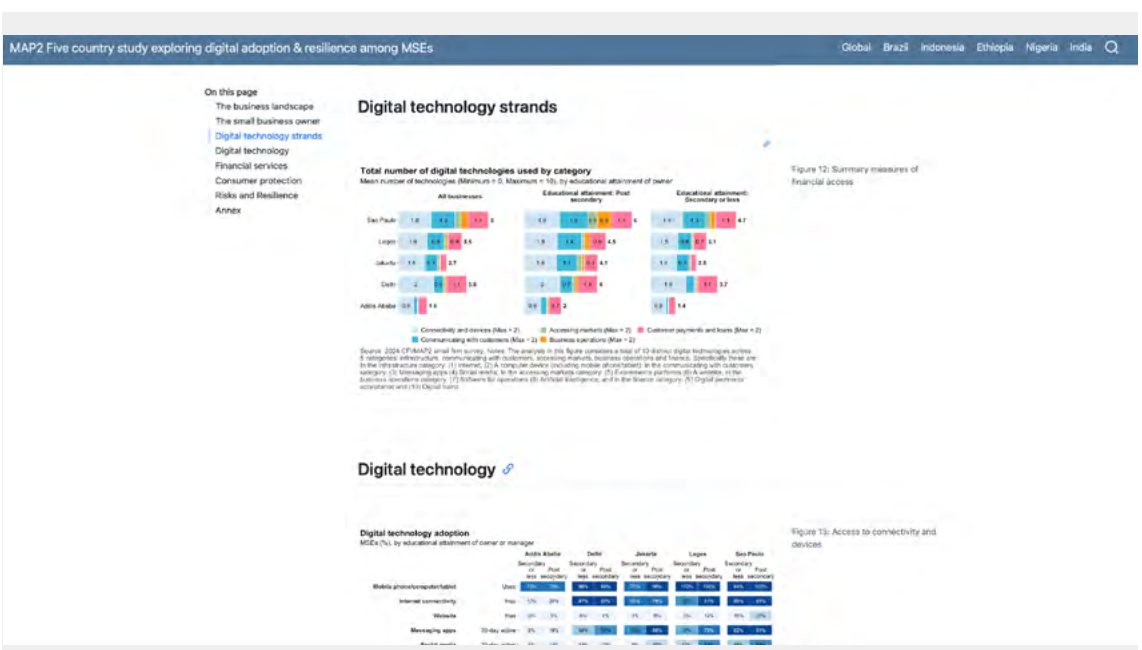
3. **Block Explorer:** https://github.com/Center-for-Financial-Inclusion/cfi-map2-blockexplorer

This repository contains R code for an R Shiny interactive web app that visualizes the geospatial sampling grids for the CFI study.



4. **Analysis and visualization:** https://github.com/Center-for-Financial-Inclusion/cfi-map2-data

This repository contains all the code for the set of analyses and figures used for the CFI study, including a GitHub pages deployment of all figures and analysis used to inform the final report: https://center-for-financial-inclusion.github.io/cfi-map2-data/cfi_map2_global.html



09 References

Aga, G., Francis, D.C., Jolevski, F., Rodriguez Meza, J. L., & Wimpey, J. S. (2023). An application of adaptive cluster sampling to surveying informal businesses. Journal of Survey Statistics and Methodology, 11(5), 1246–1266. https://doi.org/10.1093/jssam/smac037

Thompson, S. K. (1990). Adaptive cluster sampling. Journal of the American Statistical Association, 85(412), 1050–1059. https://doi.org/10.1080/01621459.1990.10474975

Totolo, E., Gubbins, P., Ruiz, L.A., Chakraborty, A., McGuinness, L., & Rice, C. (2025). Small Firms, Big Impact: Digitization, Financial Services, and Climate Resilience in Five Emerging Markets. Center for Financial Inclusion. https://www. centerforfinancialinclusion.org/wp-content/uploads/2025/08/CFI_Too-Small-to-Cover_Report_August2025.pdf

World Bank Enterprise Analysis Unit. (2016 – present). Informal sector enterprise surveys [Data set]. World Bank. https://www.enterprisesurveys.org/en/data

CFI. (2024). Project documentation: sampling methodology notes for Brazil, Ethiopia, India, Indonesia, and Nigeria; ACS weighting manual; enumerator training materials.

CENTER *for* FINANCIAL INCLUSION | ACCION

The Center for Financial Inclusion (CFI) is an independent think tank housed at Accion. CFI works to advance inclusive financial services for the billions of people who currently lack the financial tools needed to improve their lives and prosper. We leverage partnerships to conduct rigorous research and test promising solutions, and then advocate for evidence-based change.

www.centerforfinancialinclusion.org

Center for Financial Inclusion (CFI)