**The Informal Sector Enterprise Surveys in Cities in Peru, 2022**

## I. Introduction

This document provides information on the Enterprise Surveys of the Informal Sector (ESIS) data collected by the World Bank Group's (WBG) Enterprise Analysis Unit (DECEA) in two cities in Peru. The survey covers the following cities: Lima and Trujillo. The fieldwork was implemented by Datum Internacional S.A., and the data was collected between June and September 2022.

The primary objectives of the ESIS are to: i) understand the demographics of the informal sector in the covered cities, ii) describe the environment within which these businesses operate, and iii) enable data analysis based on the samples that are representative at each city level.

This report briefly describes the sampling design of the data, the structure of the dataset as well as additional information that may be useful for data users.

## II. Definition of Informality and Universe

The ESIS cover all informal businesses within a well-defined geographic area, typically a city. All eligible businesses are considered as forming the universe of inference for the survey. Informality is defined as all businesses that are not legally registered with the government and, therefore, are excluded from taxation. The exact details of these criteria vary by country. For the 2022 ESIS in cities in Peru, a business is considered informal if it is not registered with SUNAT (Superintendencia Nacional de Aduanas y de Administración Tributaria). Thus, the universe of the survey includes all businesses that meet the above definition of informality, including all sectors of activity of any size. The universe excludes, however, any illicit or illegal activity.

## III. Sampling Approach

This section briefly outlines the sampling approach of the ESIS (please see Aga et al. (2022) for further details). A challenge to conducting a representative survey of informal sector businesses is the lack of a proper sampling frame since these businesses are not registered and, therefore, they are almost always absent from official registries or any other potential sampling frame. The ESIS follow an area-based sampling methodology. Each city is overlayed with a spatial grid, dividing the area into squares of equal size.[1] In the case of the 2022 ESIS in cities in Peru these squares measured 150 meters by 150 meters. This spatial grid is prepared using a GIS software. Often, the boundaries of the city match a set of administrative boundaries. Nevertheless, when appropriate, a wider urban area (possibly outside of administrative boundaries) is used to best capture the extent of informal business activity in an area. The resulting map is reviewed and approved by the relevant WBG Task Team Leader (TTL) in consultation with the implementing contractor. For the purposes of the 2022 ISES in Peru, the shapefiles outlining administrative

[1] This methodology can be applied to any geographic area. The Informal Sector Enterprise Surveys are typically conducted within a city/well-defined urban area.

boundaries were obtained from the United Nations Office for the Coordination of Humanitarian Affairs website (https://data.humdata.org/dataset/cod-ab-per). These outlines are given in Appendix A (dashed lines on the maps), along with the spatial grid used for each city.

Each square in the spatial grid, referred to as a block, constitutes a Primary Sampling Unit (PSU). The ESIS use the Adaptive Cluster Sampling (ACS) method (Thompson, 1990, 1991), whereby an initial sample of starting blocks ($n$) is randomly selected from the city's grid. See Appendix B for the number of starting blocks selected for each city. All encountered informal businesses within each starting block are then fully enumerated. ACS takes advantage of the expectation that informal businesses are geographically clustered. To do so, a threshold number of informal businesses is defined. All blocks that meet the threshold trigger full business enumeration of all surrounding blocks. This process of sample expansion continues until no blocks meet the set expansion threshold.[2]

Each sampled block (starting or selected through subsequent expansions) is thoroughly enumerated whereby basic information (such as sector, size, etc.) for all informal businesses located in the block is collected (listed businesses). Enumerators are instructed to list and account for businesses that refuse the listing exercise as well as those that are unavailable at the time of fieldwork (e.g., they are encountered during off-hours). In those cases, enumerators record the refusal/unavailability and record a few pieces of information by observation. A randomly selected subset of the available enumerated businesses is interviewed through a 25-minute questionnaire ("interview"). As such, the ESIS s follow a two-stage sampling process, whereby a probability of selection is clearly defined by integrating selection probabilities from both stages for each interviewed business resulting in a sample that is representative of the informal sector at the city level.

For increased precision, the ESIS use the stratified version of the ACS. After the spatial grid is created, blocks in the grid of each city are categorized into four strata: central business districts, residential, agricultural, and market centers.[3] The stratification was based on local knowledge of cities by the respective field teams of Datum Internacional S.A. Note that blocks stratified as market centers are not selected using ACS but rather through simple random sampling (SRS).

## IV. Survey Implementation

This survey was fully implemented using the World Banks' Survey Solutions CAPI software. The selection for interviews was conducted in the field and in real time, via a random algorithm programmed in the CAPI so that enumerators do not have control over who gets selected. A detailed monitoring protocol was put in place during the data collection to ensure the integrity of the fieldwork and methodology. In addition to supervision through assigned supervisors, every enumerator records his/her path using a tracking software (Oruxmaps) installed

[2] Different patterns of expansion can be implemented through this methodology. The Informal Sector Enterprise Surveys typically activate all surrounding blocks of those meeting the expansion threshold. If the expansion process is considerably longer than expected, a TTL can force a stop to expansions.

[3] There is a fifth category for blocks that are physically inaccessible. This category is excluded from the sampling.

on all CAPI tablets. Enumerators submit captured paths to a centralized server at the end of the enumeration of every block. This tracking path is checked to ensure that enumerators have fully covered the block assigned to them.[4] This quality check was done at high frequency. In cases where the tracking path indicated a lower than acceptable level of effort in listing informal business, the enumerator was asked to re-survey the block. For detailed information about the total number of blocks covered, informal businesses enumerated, and interviews conducted in each city, please see Appendix B.

## V. Database Structure

The main datafile is based on interviews conducted through a standardized questionnaire, which was piloted and reviewed before fieldwork, building on previous modules used by the World Bank's Enterprise Analysis Unit to survey informal businesses. For more information about the questionnaire, see Aberra et al. (2022). Together with the standardized questions, the 2022 Informal Sector Enterprise Surveys in cities in Peru questionnaire includes a set of Peru-specific questions, all of which are marked with prefix "PER". For more details, please see the questionnaire provided as part of the survey documentation. The language in which the interview was conducted is given in variable a1a.[5]

The data contains a unique business identifier variable named *idstd*. Most variables in the dataset are numeric, except those with suffix “x”, which are string variables. The variables *wweak, wmedian, wstrict* are the (consolidated from the two stages) sampling weights corresponding to the different weighting criteria (see next section for details). The sampling weights enable inferences to the population of informal businesses in each city and are strongly recommended to be used in data analysis. The variable *strata* indicates the sampling design stratum within each city. The variable *id_square* identifies businesses that have been interviewed in the same block.

All monetary questions record the answers in the local currency unit, Peruvian Sol.

## VI. Sampling Weights[6]

To estimate population parameters, weights are applied to survey samples. In survey designs following standard SRS, the selection probability of all units is known before the actual data collection. Hence, weights can be derived as the inverse of the probability of selection. A similar process is followed for ACS, with the probability of selection (within stratum) adjusted to account for the adaptive selection process.

Let $n_h$ denote the total number of starting blocks (PSUs) selected by stratum $h$, with $h = 1, \ldots, H$ indexing each stratum. Let $N_h$ denote the total number of blocks in the universe of stratum $h$. Within stratum, starting PSUs are selected using SRS without replacement. Whenever the number of informal businesses in a given starting block is above a pre-defined threshold,

[4] Oruxmaps captures not only the path, but also how long an enumerator stayed in a block, the pace at which s/he is travelling through it, etc.
[5] The interviews were conducted in Spanish.
[6] For seminal discussions of adaptive cluster sampling, including issue of sampling weights and proper estimators to use, see Thompson (1990, 1991). Discussions and notation in this section draws heavily, among others, on Thompson (2012), Turk and Borkowski (2005), and Tout (2009).

determined by city, all surrounding blocks are enumerated as long as they are not inaccessible or market centers (market centers are selected by SRS and are not eligible for expansion). The enumeration of surrounding blocks continues until no block meets the pre-defined threshold requirement. This process produces a set of *networks*, which are formed by a group of contiguous blocks that all meet the expansion threshold. Thompson (1990, 1991) defines a network as any group of blocks where the enumeration of one block would lead to the enumeration of all blocks in the network. This definition helps show how, within a network, the probability of selection can be estimated. Specifically, let $m_i$ denote the total number of blocks in a network $i$.[7] In the case that a starting block $i$ does not meet the threshold, it is considered as a network with a size of $m_i = 1$. Then, the inclusion probability of a block in network $i$ is defined as $\pi_i$ and given by[8]:

$$\pi_i = 1 - \prod_{h=1}^{H} \left[ \frac{\binom{N_h - m_{h,i}}{n_h}}{\binom{N_h}{n_h}} \right]$$

The inverse of $\pi_i$ provides the first-stage weight for each block belonging to network $i$. For market center blocks, the weight is calculated as $\frac{n_h}{N_h}$.

In providing population estimates using ACS (with stratification), Thompson (1990, 1991) has shown that these weights produce unbiased estimators; however, in many cases there will be informal businesses that are enumerated in blocks that are not part of the starting selection and do not meet the expansion threshold. These blocks are called *edge units*, and the unbiased estimators call for omitting these blocks. The reason for this is that for edge units, there are neighboring, surrounding blocks that have not been enumerated, and so the actual probability of selection for edge units cannot be known. Due to the expense of fieldwork and the fact that there are often interviews obtained in edge units, the choice has been made to calculate weights for edge units as well. These weights are estimated according to the formula for $\pi_i$ with the following substitutions: the term $m_{h,i}$ is replaced with $(2m_{h,i} + e_h)$ where $e_h$ is a dummy that equals 1 for edge units of strata $h$ and 0 otherwise, and $n_h$ is replaced with $2n_h$.[9] The resulting weights have a known but directionally agnostic bias on population estimates when edge units are retained in this manner. As of the publication of 2022 Informal Sector Enterprise Surveys in cities in Peru, we feel it is worth the bias to retain the information collected.

The ESIS datasets are also published with variables that identify various groupings of blocks: *M_ID* identifies the network to which a block belongs. *C_ID* denotes what is termed a *cluster*, which is a contiguous group of enumerated blocks, regardless of whether those blocks are from the initial selection, part of expansions, and whether or not they meet the expansion threshold.[10] Data users are provided these variables for their own discretion in the treatment of

[7] Note that within a network, all blocks are treated as the same, and so some ACS literature indexes by network rather than block.
[8] If there are no blocks with stratum $h$ in the network $i$, then $N_h$ and $n_h$ are set to 0. Note that with networks of size $m_i = 1$, all blocks will be within one stratum and the probability of selection simplifies to the same probability of the initial SRS selection (within the stratum).
[9] In the special scenario where an edge unit is of a stratum that the adjacent network does not contain, the formula uses $N_h$ and $n_h$ of stratum $h$ (i.e., avoiding adjustment detailed in footnote 4 above).
[10] It is therefore possible for a cluster to contain multiple networks.

standard errors: if a user believes that they need to adjust standard errors to account for correlation within a given area, for example, they can choose to apply a clustering option by *C_ID*.

The published ESIS datasets consist of all interviews, respondents of which are selected randomly within blocks. These datasets therefore include second-stage weights that are applied to these interviews. These weights are produced by multiplying the first-stage (block-level) weight by an adjustment factor, which is given by the inverse of the ratio of the number of interviews completed to the total number of informal businesses found in the block.[11] Since within blocks, there typically are both refusals to the enumeration exercise and informal businesses that cannot be reached (and are recorded by observation), assumptions may be needed to calculate this adjustment factor. Specifically, three versions of the second-stage weights are provided based on different assumptions that are used in estimating the total number of informal businesses found in a block, with varying ways of handling refusals and unavailable businesses, as follows:

| **Assumption** | **Variable Name** | **Condition of Inclusion** |
|---|---|---|
| Strict | wstrict | All confirmed informal businesses only |
| Median | wmedian | All confirmed informal businesses and refusals that **don't** have signage/permits on display |
| Weak | wweak | All confirmed informal businesses and all refusals |

The strict assumption calculates the total number of informal businesses found in a block by including only businesses that are confirmed to be informal. The median assumption assumes that businesses that refused to participate in the survey and do not have signage and permits on display are part of the informal sector of the economy; these businesses are added to those confirmed to be informal. The weak assumption treats all refusals, regardless of signage and permits, as part of the informal economy. Therefore, by definition:

$$wweak \geq wmedian \geq wstrict$$

Data users can choose to apply whichever weight they conclude is appropriate under these assumptions. All indicators and analyses conducted by the Enterprise Survey team use the wmedian weights.

Data users should note that there is a debate on whether and how to use weights in regressions (see Deaton, 1997, pp.67; Haider et al. 2013; Lohr, 1999, chapter 11, Cochran, 1977, pp.150). There is no strong, large-sample econometric argument in favor of using weighted estimation for a common population coefficient if the underlying model varies per stratum (stratum-specific coefficient): both simple OLS and weighted OLS are inconsistent under regular conditions. However, weighted OLS has the advantage of providing an estimate that is independent of the sample design. More generally, if the regressions are descriptive of the population then

[11] Weights of the blocks that contain no completed interviews are proportionally transferred to blocks that (i) have at least one completed interview, (ii) have the same stratum, and (iii) have the same first-stage weight. If matching with these three parameters is not possible, the requirement of the same first-stage weight is replaced with a requirement that the blocks are in the same cluster. If matching is still not possible, only the first two parameters are used.

weights should be used. If the models are developed as structural relationships or behavioral models that may vary for different parts of the population, then there is no reason to use weights.

## VII. Caveats

Although all possible efforts were exerted to successfully implement this methodology, some informal business units are bound to be missed during enumeration, particularly the type of activities that are hidden on purpose. This is more likely to be the case for household-based activities, although the enumeration process involved, to the extent possible, knocking on every house in each enumerated block to check for informal business activities. Additionally, as explained above, the survey is representative only of informal businesses in the covered geographic area and not necessarily of the entire province or country.

## References


Aberra, A., Aga, G., Jolevski, F., and Karalashvili, N. (2022). "Understanding Informality: Comprehensive Business-level Data and Descriptive Findings". Forthcoming in Working Paper Series, Washington, D.C. : World Bank Group.

Aga, G., Francis D.C., Wild M. (2018) "Surveying Informal Enterprises: Applying Stratified Adaptive Cluster Sampling using CAPI with Implementation and Monitoring Tools", Draft Mimeo

Aga, G., Francis, D., Jolevski, F., Rodriguez Meza, J., and Wimpey, J. (2022). "Surveying Informal Businesses: Methodology and Applications". Policy Research working paper, no. WPS 9905 Washington, D.C. : World Bank Group.

Cochran, William G., Sampling Techniques, 1977.

Deaton, A. (1997) The analysis of household surveys: A Microeconometric approach to development policy, Johns Hopkins University Press, Baltimore, MD.

Greig-Smith, P. (1964) Quantitative plant ecology. (2nd ed.) Butterworths, London.

Haider, S., Solon, G., and Wooldridge, G. (2013) "What Are We waiting for?", NBER Working Paper 18859.

Jolevski, F., Aga, G. (2019) "Shedding light on the informal economy: A different methodology and new data." Let's Talk Development.

Levy, Paul S. and Stanley Lemeshow, Sampling of Populations: Methods and Applications, 1999.

Lohr, Sharon L. Sampling: Design and Techniques, 1999.

Scheaffer, Richard L.; Mendenhall, W.; Lyman, R., Elementary Survey Sampling, Fifth Edition, 1996.

Thompson, S. K. (1990). Adaptive cluster sampling. Journal of the American Statistical Association, 85(412), 1050-1059.

Thompson, S. K. (1991). Stratified adaptive cluster sampling. Biometrika, 78(2), 389-397.

## Appendix A: Primary Sampling Units, Sampling Frames

### Figure A-1: Lima

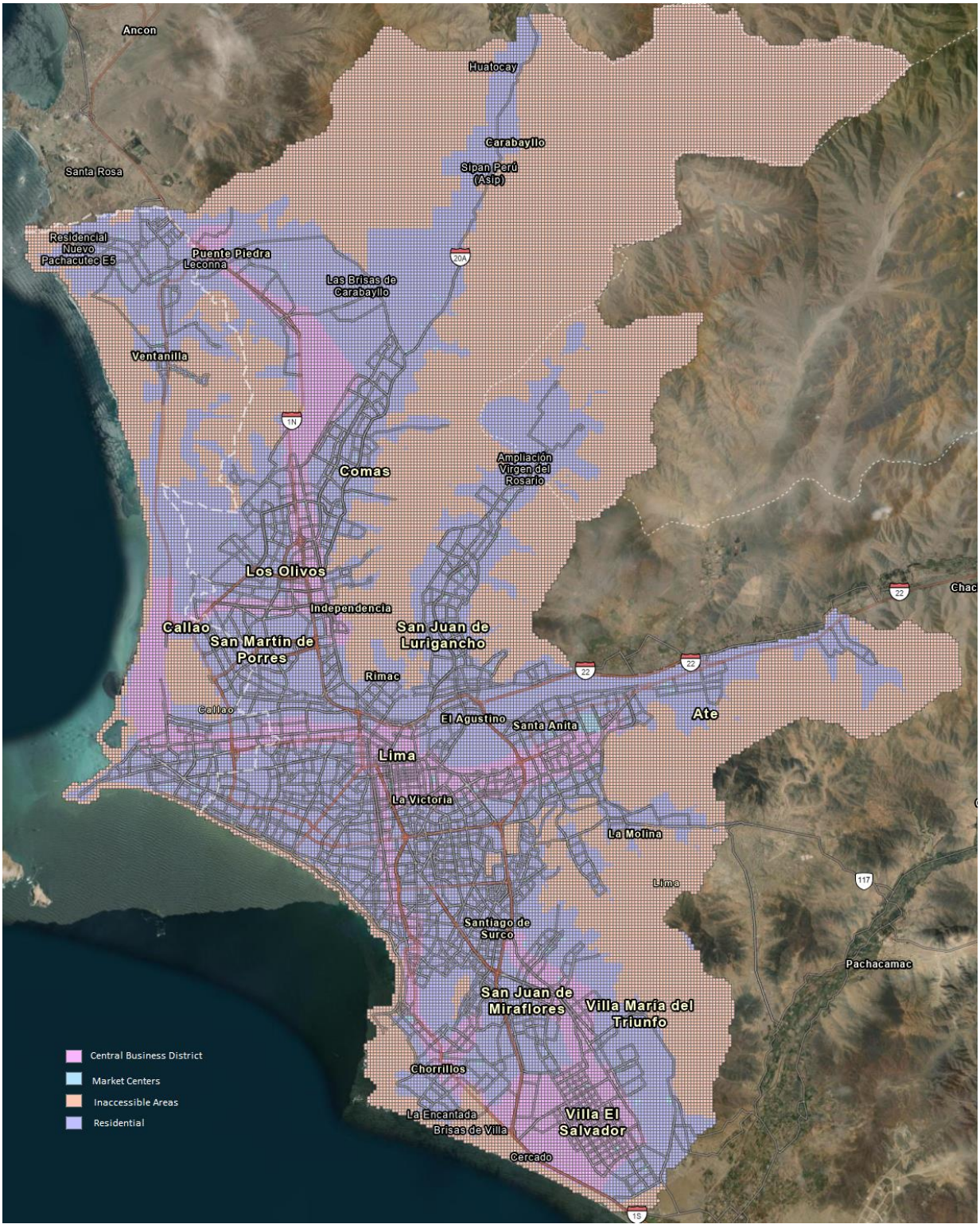

**Figure A-2: Trujillo**

**Appendix B: Details of Survey Implementation**

**Table B-1: Number of blocks enumerated, and interviews completed**

| City | Strata | Universe of blocks | Starting blocks enumerated | Total number of blocks enumerated | Informal business units enumerated | Average number of informal business units per block | Interviews completed |
|---|---|---|---|---|---|---|---|
| **Lima** | Central Business Districts | 6459 | 162 | 263 | 813 | 3.09 | 220 |
| | Residential | 23755 | 476 | 858 | 2425 | 2.83 | 574 |
| | Market centers | 364 | 193 | 193 | 653 | 3.38 | 216 |
| | TOTAL | 30578 | 831 | 1314 | 3891 | | 1010 |
| **Trujillo** | Central Business Districts | 672 | 81 | 148 | 400 | 2.70 | 99 |
| | Residential | 3559 | 250 | 599 | 1352 | 2.26 | 360 |
| | Agricultural | 8009 | 41 | 62 | 6 | 0.10 | 0 |
| | Market centers | 201 | 81 | 81 | 491 | 6.06 | 171 |
| | TOTAL | 12441 | 453 | 890 | 2249 | | 630 |
| | **GRAND TOTAL** | **43019** | **1284** | **2204** | **6140** | | **1640** |

**Table B-2: ACS parameters, networks, and clusters**

| City | Threshold for ACS expansions | Number of networks enumerated | Number of clusters enumerated |
|---|---|---|---|
| Lima | 8 | 830 | 685 |
| Trujillo | 5 | 446 | 261 |