The 2019 Zambia Micro-enterprises Survey Dataset

I. Introduction

This document provides additional information on World Bank Group (WBG) 2019 Zambia Micro-enterprises Survey data collected by the Enterprise Analysis Unit (DECEA) in Zambia. The survey covers three cities: Kitwe, Lusaka and Ndola. The fieldwork was implemented by Ipsos Zambia, a survey firm based in Lusaka, and the data was collected over the months of August 2019 to February 2020.

The primary objectives of the survey are: i) to understand the business demographics of micro-enterprises in the three cities, and ii) to describe the environment within which these businesses operate. A secondary objective of the survey is to provide an estimate of the number of micro-enterprises operating in these cities.

The report outlines and describes the sampling design of the data, the structure of dataset as well as additional information that may be useful when using the data.

II. Universe and Definition of Micro-enterprises

The universe includes formally registered businesses with less than five employees. The definition of formal registration can vary by country. For the survey in Zambia, a business that falls under any of the following two categories, in terms of registration, is considered as micro-enterprises: i) registered with Zambia Revenue Authority (ZRA) and has less than 5 employees; and ii) registered with Zambia's Patents and Companies Registration Agency (PACRA) and has council permit, but not registered with ZRA.

In terms of sector and size, the survey covers all non-agricultural sectors and businesses of all size categories if they meet the registration and size criteria.

III. Sampling Approach

A challenge to conducting a representative sample survey of Micro-enterprises is the lack of a comprehensive sampling frame of establishments. The 2019 Zambia micro-enterprises survey uses an

innovative technique[1] to survey these businesses, used by the WBG Enterprise Analysis Unit to survey informal businesses. The survey follows an area-based[2] sampling methodology with geographic area rather than an establishment or a business unit as a primary sampling unit. To account for potential clustering of these business, the survey uses an area-based sampling called (stratified) Adaptive Cluster Sampling (ACS)[3], whereby one selects a sample of starting squares and adaptively samples surrounding squares based on the number of businesses discovered in the enumerated squares. All business in selected squares will be enumerated using a 2 to 3-minutes questionnaire, referred to in this document as the short-form questionnaire. The short form questionnaire is a listing questionnaire where basic information about the business is collected. A randomly selected subset of the enumerated businesses will be given a 20-minutes questionnaire, referred to in this document as the long-form questionnaire. This is the main questionnaire of the survey and the basis of the database posted on the ES portal.

The survey is adaptive in the sense that if the number of businesses units in a square exceeds a predefined threshold, all the squares surrounding the starting square are surveyed, following the same approach of enumeration and randomly conducting the main interview. If one of the surrounding squares exceed the threshold, then the squares surrounding that square in turn are also surveyed. This process continues until either the network is exhausted, or an arbitrary cut-off point is defined.

The first step in the sampling approach is the construction of a spatial grid as the Primary Sampling Units (PSU) frame, as shown in Appendix A - 1 for Kitwe, 2 Lusaka and 3 for Ndola respectively. The grid covered the total of municipal areas and each cell had a size of 150 by 150 meters. This produced a total of about **45,530** squares between the three cities, excluding squares that are considered inaccessible. The second step was to stratify each grid, with in each city, based on land use type. The grids were categorized into five strata: residential, commercial/industrial, mixed (commercial and residential), Market centres and open area.[4] The stratification was based on local knowledge of the survey implementing contractor with approval from the WBG task team leader. The third step in the sampling process was to select a pre-defined number of starting squares from each stratum for enumeration and main data collection (see Appendix B for the number of starting squares selected for each city).

IV. Survey Implementations

Enumerators were assigned to starting squares, enumerating all business units in selected squares and administering the main questionnaire to a randomly selected subset of the enumerated businesses. This survey was fully implemented using the World Banks' Survey Solutions CAPI system. The selection

[1] For more details on the World Bank Enterprise Surveys work on applying this new methodology on informal businesses, see Jolevski, F. and Aga, G. (2019) Shedding light on the informal economy: A different methodology and new data.

[2] Area frames are well known from household surveys, though they commonly use administrative boundaries as the delimitation. However, the use of a regular grid as an area frame has a long tradition in Ecological surveys (Greig-Smith 1964), although its application to business populations is relatively rare.

[3] For further detail on ACS, please see: Thompson, S. K. (1990). Adaptive cluster sampling. Journal of the American Statistical Association, 85(412), 1050-1059.

[4] Note that there is a fifth category for squares that are inaccessible. This category is excluded from the sampling.

for long-form (main) questionnaire was conducted in real time (i.e., concurrently with the listing process) using the CAPI system with a random decaying probability of selection; these minimizes issues stemming from the transitory nature of some of these businesses. An important feature of the implementation is that enumerators did not have control over who gets selected for an interview with the long-form (main) questionnaire since the CAPI does so randomly. All respondents that were not selected for the long-form were given a short-form questionnaire, which captured information on the type of activity, physical location, and the number of workers. Outright refusals were also recorded, using enumerator observation of the activity, workers observed, and whether the business had any sign and permits on display.

Overall, the enumeration started with a total of 1327 starting squares in the three cities combined, and a total of 1932 squares were enumerated in the end (see Appendix-B for detail). Out of a total of 1,784 micro-enterprises enumerated, about 97 were randomly selected and responded to the main questionnaire (i.e., the long-form), which is the main data file.

Implementation of the actual fieldwork can be daunting given the complicated nature of the sampling methodology. An intensive and extended training and piloting sessions were conducted before the launch of the fieldwork. A three-day intensive training of enumerators and field management team took place followed by two days of piloting in each of the three cities. Based on feedback from this trainings and piloting, necessary changes were made to the questionnaire and CAPI script.

A detailed monitoring protocol was put in place during the data collection phase 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 on to all the CAPI tablets. Enumerators submit captured paths to a centralized server at the end of enumeration of every square. This tracking path is checked to ensure that enumerators have fully covered the square assigned to them.[5] This quality check was done daily, and for cases where the tracking path indicated below acceptable level of effort in listing business, the enumerator was asked to re-survey the square.

V. Database Structure

The main data file is collected using a standardized questionnaire, i.e., the long-form questionnaire. The questionnaire was developed building on previous modules used by the Enterprise Analysis Unit of the World Bank to survey informal businesses and micro-enterprises.

The data contains a unique business identifier, variable name *id*. All variables are named using, first, the letter of each section and, second, the number of the variable within the section, i.e. *a1* denotes section *A*, question *1* (some exceptions apply). All variables are numeric with the exception of those variables with an "x" at the end of their names. The suffix "x" denotes that the variable is alpha-numeric. The variables *wweak, wmedian, wstrict* are the sampling weights for the main questionnaire corresponding to the different criteria (see next section for detail on the weight computation). In the traditional sense of sampling weights, they represent inflation factors to make inferences to the

[5] The software captures more than just the path, but also how long an enumerator stayed in a square, the pace at which s/he is travelling through the square etc.

population of micro-enterprises in each city. The variable *strata* is defined as a different combination of city and the stratification variable sorting squares in to low, medium and high probability squares. Users can use variable *id_cluster* for further clustering the standard error at a relatively disaggregated grouping (see section V below for definition of cluster). The variable *id_square* identifies firms that have been interviewed in the same square.

All financial questions record the answers in the local currency unit, in this case Zambian Kwacha.

VI. Sampling Weight[6]

To estimate population parameters, weights are applied to survey samples. In surveys design following standard random sampling, selection probability of all units is known before the actual data collection. Hence, weights can be derived as the inverse of selection probability.

Computation of sampling weights is a bit involved for Adaptive Cluster Sampling (ACS) since final sample size is not known *a priori*. In ACS, selection probabilities are not known a priori since sampling squares are adaptively added to the sample depending on the number of micro-enterprises found in a square. In adaptive sampling, one instead talks about empirically derived inclusion probabilities.

Let *n* denote the total number of squares selected initially, called the staring squares. Note that, in this survey, these initial sample are selected randomly without replacement. Whenever the number of businesses in a given starting square is above a pre-defined threshold all surrounding squares are enumerated as long as they are not inaccessible or market centers. The enumeration of surrounding squares continues until no square meets the pre-defined threshold requirement. This process produces set of *clusters*, which constitutes all the neighboring squares with the number of businesses above the threshold and those with below the threshold. The latter set of squares are defined as *edge units*, because these are where the expansion process essentially stops. A subset of squares in a *cluster* that meet the expansion condition are called *network*[7]. Networks can be of different sizes (i.e., the number of squares it includes). Denoted by m_i the total number of squares in a network to which square *i* belongs; the simplest network has only a single square, where $m_i = 1$. And let a_i denote the total number of squares in a network to which square *i* is an edge unit; $a_i = 0$ if a square meets the expansion threshold. Inclusion probability π_i is defined as follows:

$$\pi_{h,i} = 1 - \left[\frac{\binom{N_h - m_{h,i} - a_{h,i}}{n_h}}{\binom{N_h}{n_h}} \right]$$

with *h* indicating the corresponding stratum.[8]

[6] Seminal discussions of adaptive cluster sampling, including issue of sampling weights and proper estimators to use, is extensively discussed in Thompson (1990, 1991). Discussions and notation in this section draws heavily, among others, on Thompson (2012); Turk and Borkowski (2005); Tout (2009).

[7] Therefore, a network is a cluster with its edge units removed (Tout 2009, pp 11)

[8] An additional adjustment may need to be made if a network crosses the stratum boundaries as well as when networks overlap; however, have not been made to the current weight.

The inverse of $\pi_{h,i}$ provides the base weight. The actual weight for micro-enterprise selected to the main questionnaire (i.e., the long form questionnaire) and included in the database is further adjusted for by the probability of selection to the long-form questionnaire. The adjustment is given by the inverse of the ratio of the number of long-form interviews completed to the total number of micro-enterprises found in a square.

Two set of assumptions on the business's eligibility criteria are used to construct sample weights. These assumptions adjust universe estimates for businesses that refuse to participate in the survey.

Assumption	Variable Name	Condition of Inclusion
Strict	wstrict	All confirmed micro-enterprises only
Weak	wweak	All confirmed micro-enterprises businesses and all refusals

The strict criteria include only businesses that are confirmed to be micro-enterprise. The weak assumption treats all refusals as part of micro-enterprises. Therefore, by definition:

$$wweak > wstrict$$

It is upon the user 's discretion to determine which weight to use.

Users should note that there is a debate as to the use of weights in regressions (see Deaton, 1997, pp.67; Haider et al 2013; Lohr, 1999, chapter 11, Cochran, 1977, pp.150). There is not 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 have the advantage of providing an estimate that is independent of the sample design. More generally, if the regressions are descriptive of the population then 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, users should exercise some caveats when using the data. Despite all concerted efforts, some 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 the selected square to check for business activities. Further, as noted above, the sampling weight reported may require some further finetuning to address cases where a network crosses more than one stratum, although this would be a minor issue in the case of Zambia survey as there are few cases of stratum crossing by networks. Users should also note that the survey is representative only of businesses in the respective cities, and not necessarily of the entire province or country.

References

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.

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

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.

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

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

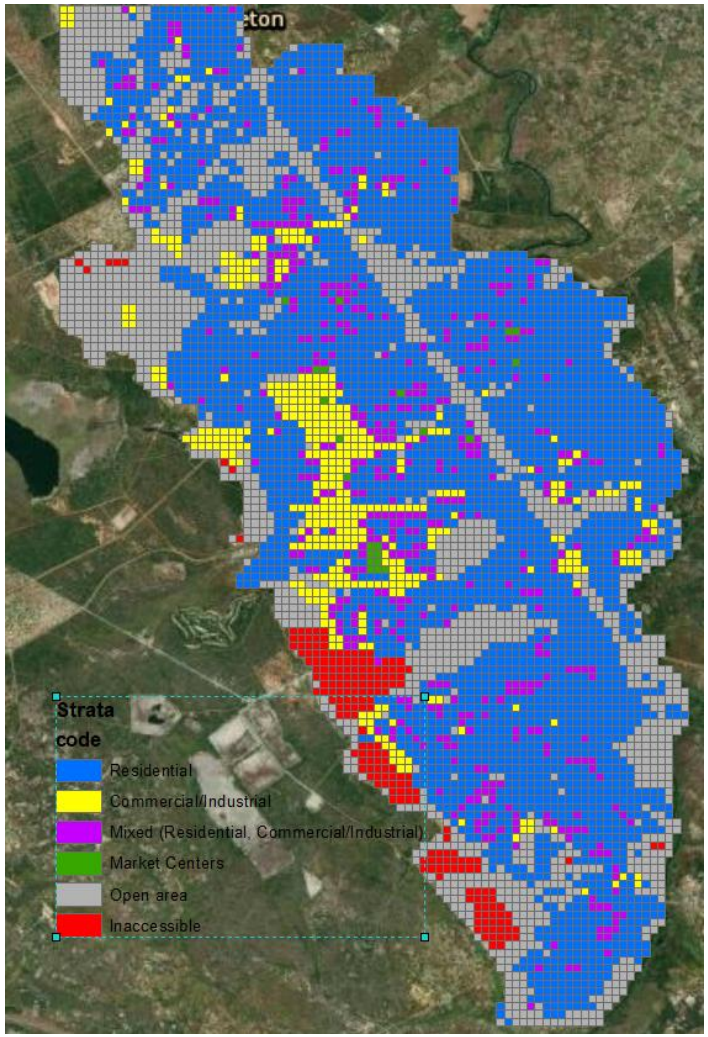
Jolevski, F., Aga, G. (2019) "Shedding light on the informal economy: A different methodology and new data." Let's Talk Development.

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

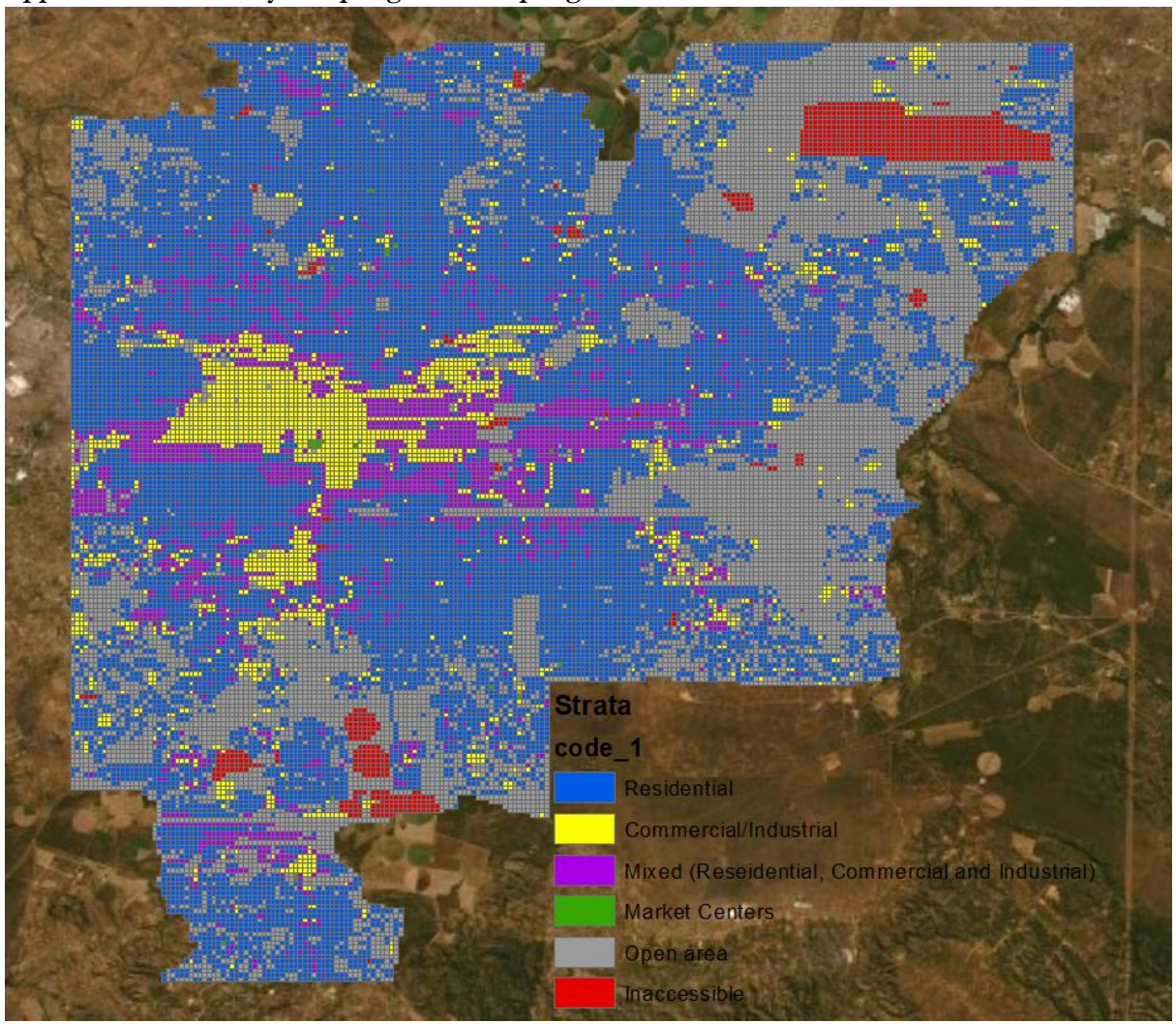
Thompson, S. K. (1990). Adaptive cluster sampling. Journal of the American Statistical Association, 85(412), 1050-1059.

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

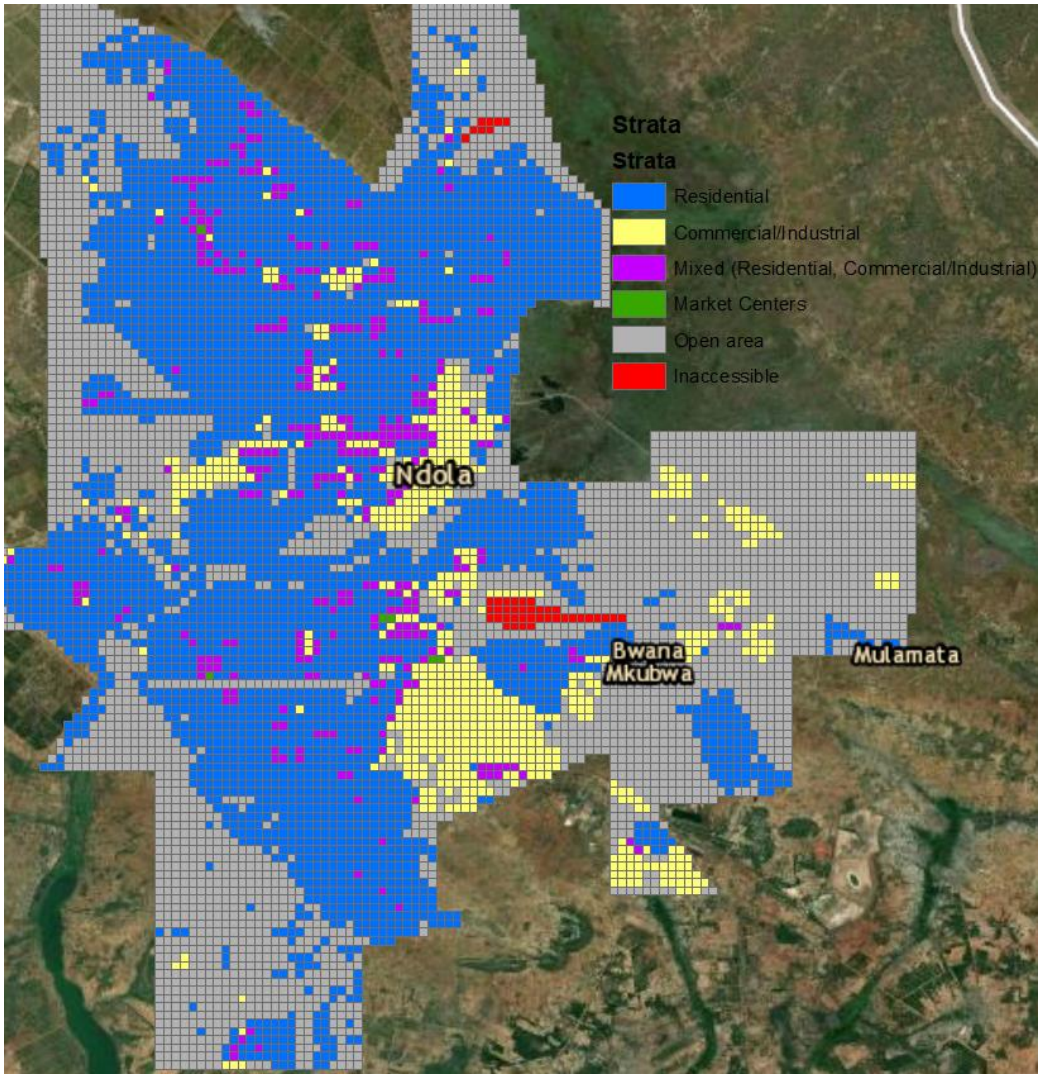
Appendix A-1: Primary Sampling Unit sampling frame for Kitwe



Appendix A-2: Primary Sampling Unit sampling frame for Lusaka



Appendix A-3: Primary Sampling Unit sampling frame for Ndola



Appendix B: Number of squares enumerated and interviews completed

City	*Starata*	*Number of Starting squares Enumerated*	*Total number of Squares Enumerated*	*Total number of micro-enterprises found*	*Average number of micro-enterprises per square*	*Total Number of long-form interviews completed*
Lusaka	Residential	395	472	315	0	53
	Commercial/Industrial	47	55	75	1	14
	Mixed	58	76	81	1	3
	Market centers	9	9	4	0	0
	Open areas	236	239	34	0	0
Kitwe	Residential	160	412	281	0	4
	Commercial/Industrial	18	34	50	1	4
	Mixed	16	67	83	1	4
	Market centers	6	6	62	10	5
	Open areas	54	59	2	0	0
Ndola	Residential	187	333	489	1	6
	Commercial/Industrial	42	39	19	0	1
	Mixed	11	39	215	5	3
	Market centers	6	6	65	10	0
	Open areas	82	86	9	0	0
	Total	**1327**	**1932**	**1784**	**2**	**97**