The Zimbabwe 2016 Informal Sector Business Survey

I. Introduction

This document provides additional information on the 2016 Informal Business Sector Survey (ISBS) data collected by the Enterprise Analysis (DECEA) in Harare, Zimbabwe, between April and July 2017. The fieldwork was implemented by Probe Market Research, a local survey firm based in Harare.

The primary objectives of the survey are: i) to understand the business demographics of the sector 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 informal businesses 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. Sampling Approach

A challenge to conducting a representative sample survey of informal sector businesses is the lack of a proper sampling frame of establishments since these businesses are not registered with government organs. Informal businesses are almost always absent from official registries and sampling frames. Further, informal businesses are usually hard to reach (often invisible on purpose), and tend to cluster in certain geographic areas, such as low income residential areas. Previous efforts to measure informal sector businesses have relied either on fully enumerated economic censuses or occurred in conjunction with household surveys. Both methods are expensive and time-intensive; while the latter method is frequently used, it does not capture informal sector units at their point of business, and it only provides a second order approximation to the population and activity of informal sector units.

The 2016 Harare ISBS pilots a new way of surveying these businesses. The survey follows an area-based[1] sampling methodology, whereby the primary sampling unit is a geographic area

[1] 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.

rather than an establishment or a business unit. To account for potential clustering of informal business, the survey uses a particular type of area-based sampling, called (stratified) Adaptive Cluster Sampling (ACS)[2]. This is a version of area-based sampling in which one selects a sample of starting grids (usually squares), which will constitute the start of the fieldwork. All informal business in selected squares will be enumerated, using a 2 to 3-minutes questionnaire (short-form questionnaire). A randomly selected subset of the enumerated businesses will be given a 20-minutes, long-form questionnaire.

The first step in the sampling approach was the construction of a spatial grid as the Primary Sampling Units (PSU) frame, as shown in Figure 1 for Harare. The grid covered the total of municipal Harare, and each cell had a size of 200 by 200 meters. This produced a total of about 22,000 grids. The second step was to stratify each grid, based on likely concentration of informal business units. In Harare, the grids were categorized into four strata: three strata of low, medium, and high concentration of informal sector activity, and a market centre. The stratification was based on local knowledge of the survey implementing contractor. The third step in the sampling process was to select a pre-defined number of starting squares from each stratum for enumeration purposes. For Harare ISBS, a total of 226 starting squares were randomly selected for enumeration (see Appendix B). The target number of starting squares, as well as the initial allocation across strata, was defined through a simulation. This simulation is implemented in R and uses the Shiny library.

III. Survey Implementation

Since the primary sampling unit is a set of grids/squares, enumerators were assigned to starting squares. All informal business units in selected squares were enumerated using a 2 to 3-minutes questionnaire, (called short-form questionnaire). A randomly selected subset of the enumerated businesses were given a 20-minutes, long-form questionnaire, essentially the main questionnaire of the survey. This survey was fully implemented into the World Banks' Survey Solutions CAPI system. The selection for long-form (man) questionnaire was conducted in real time using the CAPI system; these minimizes the issued stemming from the transitory nature of many informal activities. 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. 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 and workers observed.

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

The survey is adaptive in the sense that if the number of informal 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. We defined this cut-off for Harare ISBS to be the 4th expansion, though it was never reached in fieldwork. Overall, the enumeration started with a total of 226 starting squares and a total of 439 squares were enumerated in the end. About 3700 informal business units were listed (see Appendix-B). Out of the 3700 informal units, about 515 were randomly selected to the main questionnaire (i.e., the long-form). The main data file therefore contains 515 observations.

Implementation of the actual fieldwork can be daunting given the complicated nature of the sampling methodology. This problem gets even more relevant if this design is implemented in a low-skill environment. A series of training and piloting sessions were conducted before the launch of the fieldwork. An initial training of enumerators and field management team took place in November 2016, followed by piloting. Based on feedback from this training and piloting, necessary changes were made to the questionnaire and CAPI script. A second round of training with the entire field team took place in March 2017. It was organized as an intensive full-day training seminar followed by piloting and de-briefing on the second day. A third and final (virtual) training was conducted in the first week of April 2017 to clear any outstanding issues and fine-tune survey instrument and data collection methodology.

The use of electronic data collection devices and monitoring tools enables the implementation of more challenging types of survey designs. 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 the CAPI tablet. Enumerators submit captured paths to a centralized server at the end of enumeration of every square. This tracking path is checked by overlaying it on mapping software to ensure that enumerators have fully covered the square assigned to them. This quality check was done daily, and for cases where the tracking path indicated below acceptable level of effort in listing informal business, the enumerator was asked to re-survey the square.

IV. Database Structure

The main data file is collected using a standardized questionnaire (the long-form questionnaire) administered to randomly selected informal business units. The questionnaire was developed building on previous modules used by the Enterprise Analysis Unit of the World Bank.

There is a unique establishment 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 variable *weight_lf* is the relevant sampling weight for the main questionnaire and hence for analysis (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 population of informal businesses in Harare.

V. Sampling Weights

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 since final sample size is not known *apriori*. In ACS, selection probabilities are not known a priori since units are adaptively added to the sample depending on the number of informal units found in a square. In adaptive sampling, one instead talks about empirically derived inclusion probabilities.

Denote by *n* 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 informal businesses in a given starting square is above a pre-defined threshold, all surrounding squares will be enumerated. 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 informal firms 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*[3]. 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 $m_i =$

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

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.[4]

The inverse of $\pi_{h,i}$ provides the base weight. The actual weight for informal firms 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 informal business found in a square.

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.

VI. 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 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 the selected square to check for informal 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 Harare survey as there are few cases of stratum crossing by networks. Users should also note that the survey is representative only of informal businesses in the respective cities, and not necessarily of the entire province or country.

[4] 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.

References:

Cochran, William G., Sampling Techniques, 1977.

Gemechu Aga, David C Francis, Michael Wild (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.

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

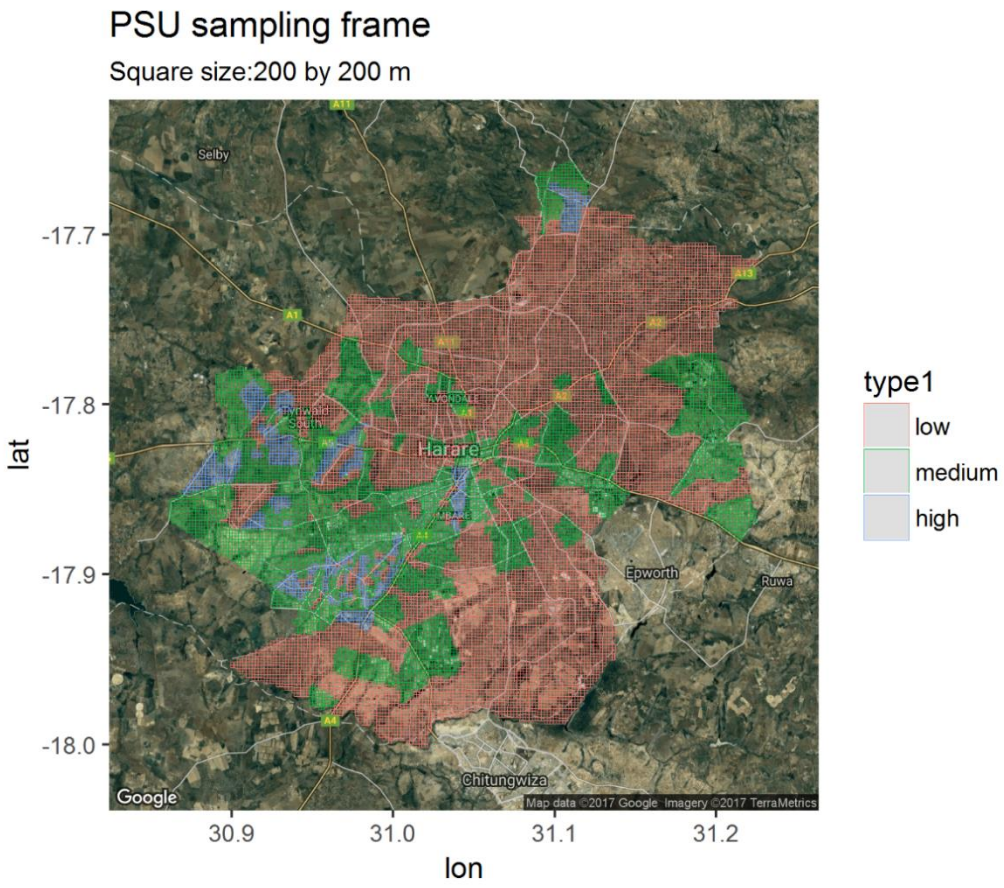
Lohr, Sharon L. Samping: Design and Techniques, 1999.

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 Unit sampling frame



Appendix B: Number of squares enumerated and informal business unit found

Starata	*Number of Starting squares Enumerated*	*Total number of Squares Enumerated*	*Total number of informal business units found*	*Average number of informal business units per square*
Low	120	145	285	2
Medium	70	146	1042	7
High	30	142	1609	11
Markets	6	6	751	125
Total	**226**	**439**	**3687**	**8**