The 2018 Mozambique Informal Sector Business Survey Dataset

I. Introduction

This document provides additional information on World Bank Group (WBG) 2018 Mozambique Informal Business Sector Survey (ISBS) data collected by the Enterprise Analysis Unit (DECEA) in Mozambique. The survey covers three cities, Beira, Maputo and Nampula, and was conducted between July and December 2018. The fieldwork was implemented by COWI Mozambique Lda, a survey firm based in Maputo.

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. Universe and Definition of Informality

The universe includes informal businesses, where informality is defined based on whether or not a business is formally registered with the government. The definition of formal registration can vary by country. For Mozambique survey, a business that lacks at least one of the following three items is considered as informal: i) operating license (e.g., from municipality, or BAU); ii) business registration certificate (e.g., Conservatória Do Registo Das Entidades Legais Or Balcão De Atendimento Único (BAU)); and iii) taxpayer's identification number (or NUIT) in the name of the owner or business.

Informality is often equated with illicit activities, but our universe excludes those types of activities. Therefore, the universe of the survey is businesses whose products or services are not illicit per the country's law but are produced by entities that are not formally registered with the government. In terms of sector and size, the survey covers all non-agricultural sectors and businesses of all size categories if they meet the informality condition.

III. 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. Consequently, 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, bus or train stations, etc. Previous efforts to measure informal sector businesses have relied either on fully enumerated economic censuses or surveys conducted 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 2018 Mozambique ISBS uses an innovative technique to survey these businesses. The survey follows an area-based[1] sampling methodology with geographic area rather than an establishment or a business unit as a primary sampling unit. To account for potential clustering of informal business, the survey uses an area-based sampling called (stratified) Adaptive Cluster Sampling (ACS)[2], whereby one selects a sample of starting squares and adaptively samples surrounding squares based on the number of informal firms discovered in the enumerated squares. All informal 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 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.

[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.

[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 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-1,2 and 3 Maputo, Beira and Nampula respectively. The grid covered the total of municipal[3] areas and each cell had a size of 150 by 150 meters. This produced a total of about 24,000 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 likely concentration of informal business units. The grids were categorized into four strata: three strata of low, medium, and high concentration of informal sector activity, and a market centre.[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 informal 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 for long-form (main) questionnaire was conducted in real time (i.e., concurrently with the listing process) using the CAPI system; these minimizes issues 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 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 and workers observed.

Overall, the enumeration started with a total of 394 starting squares in the three cities combined, and a total of 982 squares were enumerated in the end. About 11,000 informal business units were listed in total (see Appendix-B for detail). Out of the 11,000, about 540 were randomly selected 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 four days intensive training of enumerators and field management team took place followed by a one-day piloting in each of the three

[3] Excludes Matola in the case of Maputo.

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

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 informal 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.

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. All variables with prefix "MZ" are Mozambique specific questions. 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 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 *clustered* for further clustering the standard error at a relatively disaggregated grouping (see section V below for definition of cluster).

[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.

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 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*[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]

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.

---

[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.

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 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 Mozambique 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.

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.

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.

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.

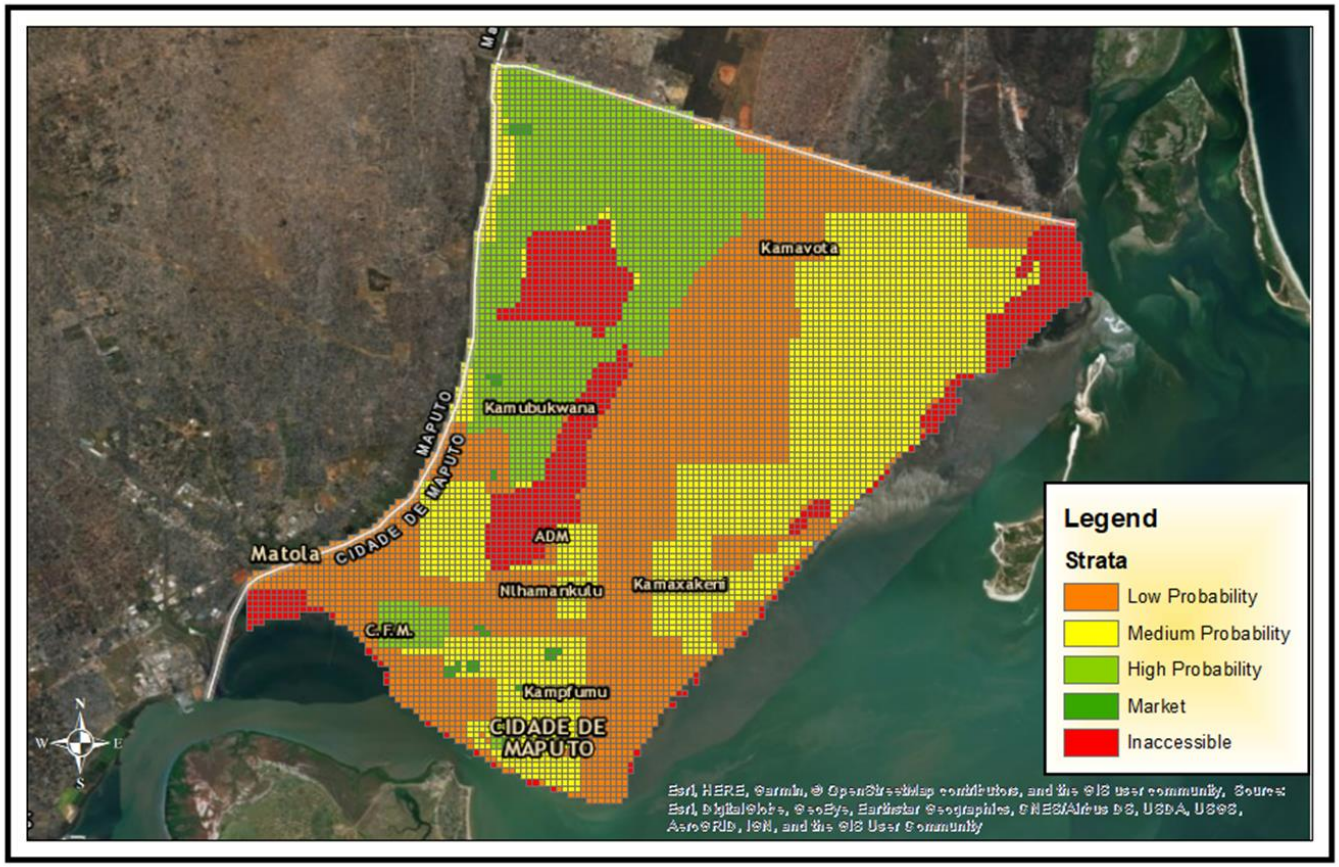
Lohr, Sharon L. Samping: Design and Techniques, 1999.

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

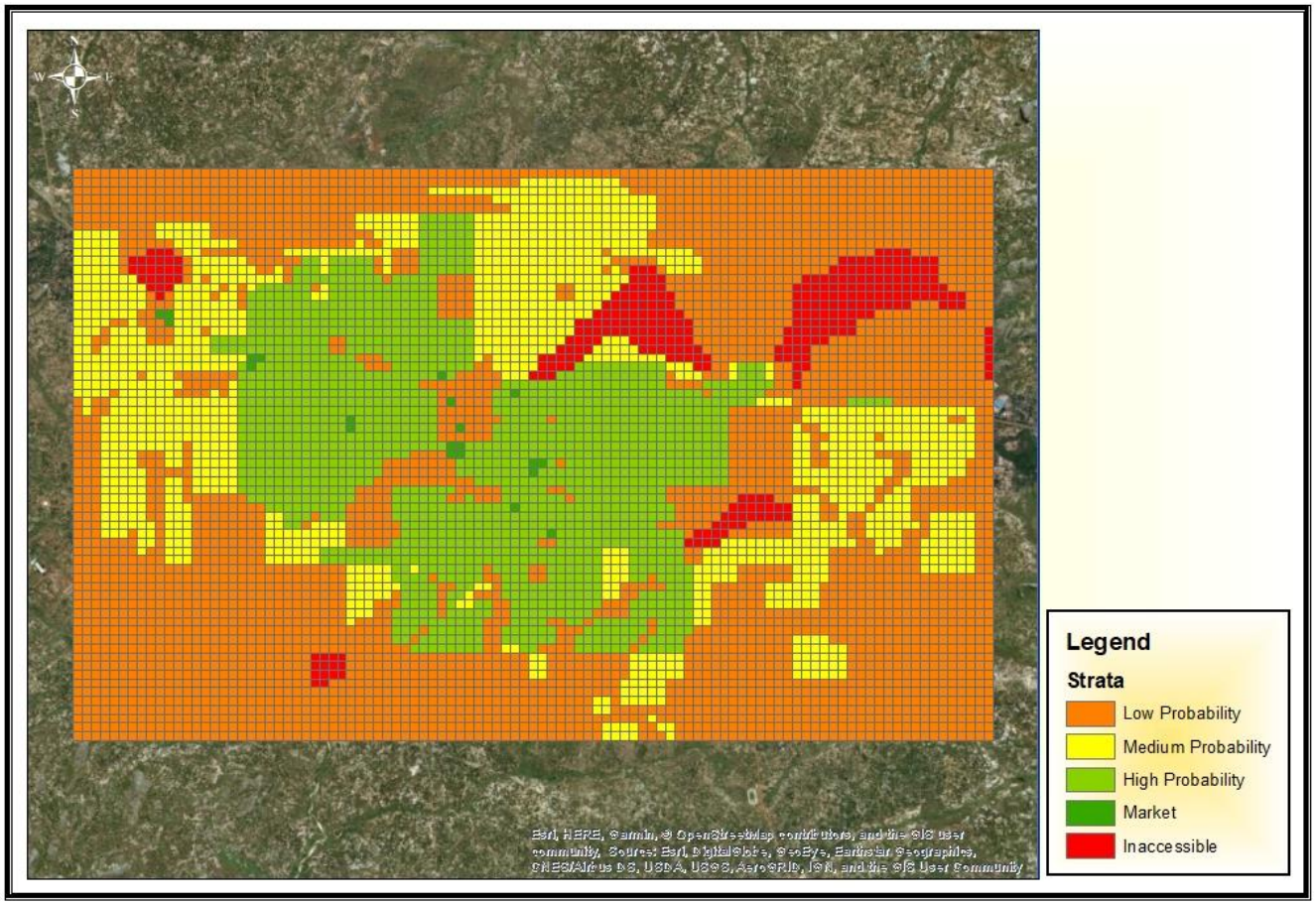
Thompson, S. K. (1990). Adaptive cluster sampling. Journal of the American Statistical Association, 85(412), 1050-1059.

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

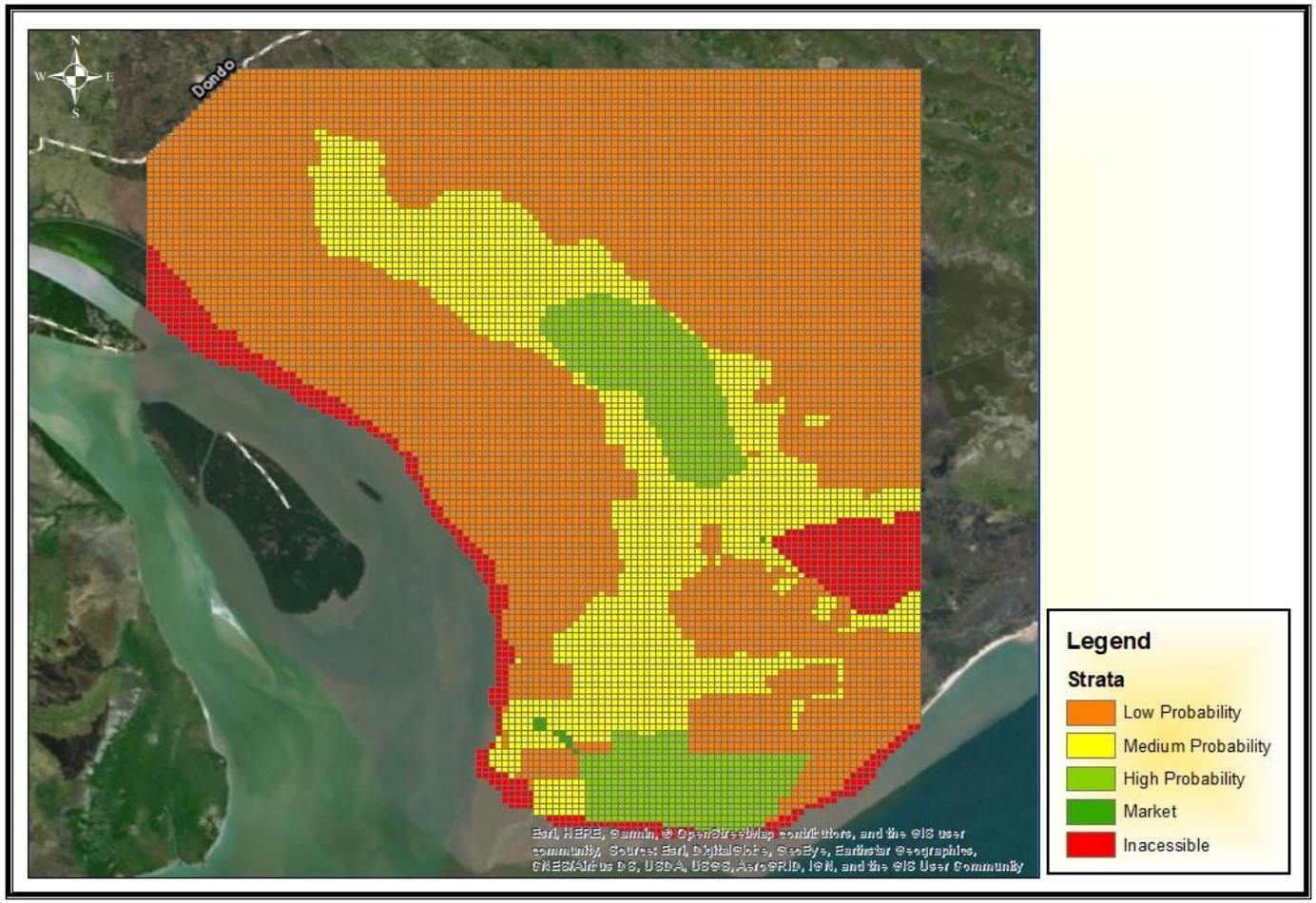
Appendix A-1: Primary Sampling Unit sampling frame for Maputo



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



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



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

City	*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*	*Total Number of long-form interviews completed*
City of Beira	Low Probability of Informality	73	91	210	2	16
	Medium Probability of Informality	35	93	649	6	63
	High Probability of Informality	19	74	766	10	77
	Market Center	3	3	156	52	6
City of Nampula	Low Probability of Informality	68	118	867	7	32
	Medium Probability of Informality	28	102	1082	10	32
	High Probability of Informality	34	127	1525	12	46
	Market Center	5	5	172	34	13
Maputo City	Low Probability of Informality	61	178	1895	10	108
	Medium Probability of Informality	39	99	1318	13	73
	High Probability of Informality	22	85	1010	11	52
	Market Center	7	7	1432	204	36
	Total	**394**	**982**	**11082**	**31**	**554**