LGPI Survey 2019

A User Guide

GLD

Produced by: Erica A. Metheney

erica.metheney@gu.se

If you have any questions contact me at erica.metheney@gu.se

Table of Contents

1 Overview **3**

1.1 Technology . . . 3

1.2 Data Collection . . . 4

1.2.1 Location . . . 4

1.2.2 Dates . . . 4

1.2.3 Survey Implementation . . . 4

2 Codebook Contents **5**

2.1 Details . . . 5

3 Variable Names **7**

3.1 Sub-Datasets . . . 7

3.1.1 Sub-dataset Codes . . . 8

3.2 Naming Scheme of Variables . . . 9

3.2.1 Multiple Choice Example . . . 10

3.2.2 Multi Select . . . 10

3.2.3 Manual Entry . . . 10

3.2.4 Questions Asked in Loops . . . 11

3.2.5 Questions Asked in Double Loops . . . 11

4 Coding of Variables **12**

4.1 Missing Data . . . 12

4.2 Don't Know/Refuse to Answer . . . 12

4.3 Numerical and Categorical Variables . . . 13

4.3.1 Numerical . . . 13

4.3.2 Categorical . . . 13

5 Importing, Encoding, and Merging Data **14**

5.1 STATA . . . 14

5.1.1 Importing . . . 14

5.1.2 Encoding of Variables . . . 14

5.1.3 Merging Datasets . . . 15

5.2 R . . . 16
5.2.1 Importing . . . 16
5.2.2 Encoding of Variables . . . 16
5.2.3 Merging Datasets . . . 16

6 Data Cleaning **17**
6.1 Dropped Observations . . . 17
6.2 Don't Know and Refuse to Answer Responses . . . 18

7 Survey Weights **19**
7.1 Sampling & Post–Stratification Weights . . . 19
7.2 Using the Weights . . . 20
7.3 Checking the Weights . . . 21

8 Sampling Plan of LGPI 2019 **22**
8.1 Overview of Sampling Plans . . . 24
8.2 Defining the Bins . . . 25
8.2.1 Creating Urban Bins . . . 25
8.2.2 Creating Rural Bins . . . 26
8.3 Selection of Square Kilometers . . . 27
8.3.1 Capital Regions . . . 27
8.3.2 Border Regions . . . 27
8.3.3 Sampled Square Kilometer Evaluation . . . 28
8.4 Selection of Hectares . . . 28
8.5 Selection of Houses Respondents . . . 28
8.6 Reaching Target Number of Observations . . . 28
8.6.1 Team Size . . . 29
8.6.2 Special Considerations . . . 29

About This File

This file was created to assist with research based on data collected from the LGPI Survey 2019.

SECTION 1

Overview

This document describes how to utilize the LGPI codebook and LGPI data for your research. Specifically you will find:

- Structure of the data
- Naming scheme of variables
- Coding methods used for different types of data
- How to import the data into both STATA and R
- How to incorporate sample weights

Refer to the table of contents for complete list of all available information.

We now present some details about the creation and implementation of LGPI 2019.

1.1 Technology

The survey was created in SurveyToGo, a product of Dooblo https://www.dooblo.net/downloads/.

The underlying code of SurveyToGo is JavaScript.

1.2 Data Collection

1.2.1 Location

The LGPI was conducted in 5 regions within 3 countries. Of the five regions, 3 are capital cities and 2 are border areas.

Countries: Kenya, Zambia, Malawi

Regions: Nairobi, Lusaka, Lilongwe, Zambia Border (with Malawi), Malawi Border (with Zambia)

1.2.2 Dates

Kenya: May 2019 – August 2019

Zambia: May 2019 – August 2019

Malawi: September 2019 – November 2019

1.2.3 Survey Implementation

Below is the list of partners who implemented the survey.

Kenya: IPSOS https://www.ipsos.com/en-ke

Zambia:

Malawi: Institute of Public Opinion and Research (IPOR) http://www.ipormw.org

SECTION 2

Codebook Contents

The codebook is available in Excel (.xlsx) format and contains the following information:

- Variable Name
- Question
- Index
- Answer Choices
- Conditions
- Question Type
- VarType General
- VarType Specific

2.1 Details

Variable Name

The variable names are determined by the scheme outlined in Section 2.

Question

Question asked in the survey.

Index

This column represents the order that the answer choices were presented in the survey. Note this information is only relevant for multiple choice/multi select questions.

Answer Choices

For multiple choice and multiple select questions the answer choices are listed in the order they were presented in the survey.

Conditions

This columns gives, in plain English, the conditions that had to be met for the question to be asked.

Question Type

In the survey there are different types of questions.

- Multiple Choice - respondents may choose exactly one option
- Multiple Select - respondents may choose as many options as they wish
- Manual Entry - answer is manually typed in by the enumerator

VarType General

This columns states whether the data collected is numerical (stored as a float) or categorical (stored as a factor

VarType Specific This column gives the specific variable type. For numerical variables this column assigns either

- Discrete - variable takes on a distinct set of values
- Continuous - variable can take values within some interval

For categorical variables this columns assigns either

- Nominal - the categories serve only as names
- Ordinal - the categories are ordered in a natural and meaningful way

IMPORTANT: In some instances the determination of VarType Specific is application dependent. When working with categorical data it is imperative to review the specific types of data you are working with in order to choose appropriate analysis methods.

SECTION 3

Variable Names

The naming scheme for variables in the LGPI dataset are based on the structure of the LGPI and the question types.

3.1 Sub-Datasets

Due to the size of the LGPI the questions of the survey are categorized into various topics, each creating their own dataset. Each sub-dataset is identified by a 4-letter code. The list of codes is given in Subsection 3.1.1.

The LGPI 2019 was conducted in 5 regions. We assign each region a 2-letter code

- KE: Nairobi,Kenya
- LU: Lusaka, Zambia
- ZB: Zambia Border (with Malawi)
- LI: Lilongwe, Malawi
- MB: Malawi Border (with Zambia)

Every sub-dataset has a name of the form XXXX_YY where XXXX is the topic code and YY is the region code. This structure allows the researcher to select those pieces which meet their research needs without downloading large amounts of unnecessary data.

3.1.1 Sub-dataset Codes

- aurc: authority, responsibility, and corruption
- basc: clothing/shelter (basic goods and services)
- brth: obtaining birth registration
- bspt: obtaining business permit
- caul: conjoint on authority and legitimacy
- ciau: community influences and authorities
- cpni: community participation, norms, and identity
- ctrl: control variables
- demographics: demographic data
- drlc: obtaining driver's license
- dthc: obtaining death certificate
- educ: education
- elpt: electoral participation
- elty: electricity
- evle: election and voting list experiment
- evlv: elections and voting levels experiment
- evmb: elections and voting mediation blunt experiment
- evmn: elections and voting mediation named experiment
- extr: extraction
- geog: geographic data
- hlth: healthcare
- kish: kish grid information
- land: land data
- lexp: land experiment

- logi: logistical data, form and subform, consent, debriefing questions
- mexp: migration experiment
- migr: migration data
- mrge: obtaining marriage certificate
- osup: other support and ability to get state service
- pnpn: party norms and participation
- pspt: obtaining a passport
- sdrs: security and dispute resolution
- sspe: small service provision experiment
- vnin: village neighborhood interaction
- vnnp: village neighborhood norms and participation
- wtsn: water and sanitation data

3.2 Naming Scheme of Variables

The structure of the variable name depends on the question type. Below we describe the structure of each variable name by variable type.

Multiple Choice	code_q#
Multi Select	code_q#_INDEX
Manual Entry	code_q#
Other Specify*	code_q#_O

*– If a question has an Other (specify) option or any answer choices that request a specified value, there will be variable containing the manually entered information with the naming convention above.

If a single question has multiple answer choices with a specify options the variable name will have the form:

$code_q\#_O$i

where i is the index of the answer choice.

Note that the last character is an upper case "O" for "Other" **not** a zero.

3.2.1 Multiple Choice Example

Question: "Is the house in an urban or rural area?"

Answer Choices: Urban, Rural

There is 1 column in the dataset that describe the data collected for this question. Suppose this was question 7 in the dataset then the variable name is land_q7 and the variables takes on the values: "Urban", "Rural", or NA

3.2.2 Multi Select

Question: "How do you know this person?"

Answer Choices: Friend, Relative, Neighbor, Work Together

There are 4 columns in the dataset that describe the data collected for this question. Suppose this was question 12 in the dataset, then the variables are

- land_q12_1 : binary variable that is equal to 1 if Friend was chosen, 0 if friend wasn't chosen, or NA if the question wasn't answered
- land_q12_2 : binary variable that is equal to 1 if Relative was chosen, 0 if friend wasn't chosen, or NA if the question wasn't answered
- land_q12_3 : binary variable that is equal to 1 if Neighbor was chosen, 0 if friend wasn't chosen, or NA if the question wasn't answered
- land_q12_4 : binary variable that is equal to 1 if Work Together was chosen, 0 if friend wasn't chosen, or NA if the question wasn't answered

3.2.3 Manual Entry

Question: "What was the name of the person?"

Answer: stated by the respondent and typed in by the enumerator

There is one column in the dataset that describes the data collected for this question. Suppose this was question 5 in the survey, then

- land_q5 : takes on the value of the string typed in by the enumerator or NA if the question was not answered

3.2.4 Questions Asked in Loops

In certain parts of the survey a set of questions is asked repeatedly for a list of individuals (ex. the same questions about school asked for every child in the household).

In these types of cases the variable names will follow the same convention as outlined above with an additional component to identify the individual.

Multiple Choice	code_q#_Ii
Multi Select	code_q#_Ii_INDEX
Manual Entry	code_q#_Ii
Other Specify*	code_q#_Ii_O

where the capital "I" stands for "Individual" and i represents the ith person in the loop.

3.2.5 Questions Asked in Double Loops

In the SDRS (security and dispute resolution) sub-dataset, there are a number of questions that were asked inside a *double* loop.

For example, we asked the respondent to choose what disputes/crimes they had experienced in the past year, and then for each dispute/crime asked them who they turned to for help. For each person they turned to for help we asked a standard set of questions. This standard set of questions exists in a double loop indexed by the dispute/crime and the person turned to for help.

In these types of cases the variable names will follow the same convention as outlined above with an additional component to identify the individual and the dispute/crime.

Multiple Choice	code_q#_Dd_Ii
Multi Select	code_q#_Dd_Ii_INDEX
Manual Entry	code_q#_Dd_Ii
Other Specify*	code_q#_Dd_Ii_O

where the capital "D"("C") stands for "Dispute"("Crime"), d represents the d (c)th dispute (crime), the capital "I" stands for "Individual", and i represents the ith person in the loop.

4

SECTION

Coding of Variables

4.1 Missing Data

All missing data is coded as NA. Note that many questions were only asked to subset of individuals so NA may represent data that is "missing" because it was not collected. Therefore when analyzing data it is imperative to consider the conditions under which the question was asked (this can be found in the codebook).

4.2 Don't Know/Refuse to Answer

Within the survey, the options Don't Know, Refuse to Answer, and Don't Know/Refuse to Answer occur frequently. The coding of these options depends on the type of variable they occur in.

Categorical Variables If there is a Don't Know, Refuse to Answer, or Don't Know/Refuse to Answer option in a categorical variable, the string "Don't Know", "Refuse to Answer", or "Don't Know/Refuse to Answer" will be used.

It is important to note that when a numerical code is assigned to "Don't Know", "Refuse to Answer", or "Don't Know/Refuse to Answer" in either STATA or R, the coding number will be equal to the place of the phrase in the answer choice list.

For instance, if "Don't Know/Refuse to Answer" is the 14th option in the answer choices, then the coding of "Don't Know/Refuse to Answer" will be 14. We emphasize this point since in many datasets this type of value would be coded as some other value such as 98 or 97. Therefore we urge the researcher to take great care in the handling of these types of values.

Numerical Variables In order to keep numerical variables coded as float data types we code these options in the following way:

- -1 = Don't Know
- -2 = Refuse to Answer
- -3 = Don't Know/Refuse to Answer

4.3 Numerical and Categorical Variables

4.3.1 Numerical

All numerical variables are coded as floats regardless of being discrete or continuous.

4.3.2 Categorical

Categorical variables are always coded as strings, not by their code value. For example if the question was "Was the person who helped you a man or a woman?" and the answer choices where Man, Woman, Don't Know/Refuse to Answer, the answer in the dataset is recorded at "Man" or "Woman" as opposed to 1 or 0. This guarantees consistent recording of data throughout the survey.

If you need assistance with encoding the string data in STATA or R please see Chapter 5.

SECTION 5

Importing, Encoding, and Merging Data

This chapter provides instructions for importing datasets, checking the encodings of variables, and merging datasets in both STATA (.dta file) and R (.rds file).

IMPORTANT: Only use the STATA file in STATA and the R file in R. Importing the data into the wrong software might result in a loss of information regarding the levels of the categorical variables.

5.1 STATA

5.1.1 Importing

- Save the STATA file (.dta) to your directory of choice.
- Click the *Open File* button in STATA and search for the file.

5.1.2 Encoding of Variables

If you would like to see all the levels of a particular variable use the command *label list VARIABLE-NAME*. See the example below.

```
. label list LAND_Q8
LAND_Q8:
           1 Village/Town Council
           2 Spouse's parents
           3 Spouse's other relatives
           4 Chief
           5 Village Head/Neighborhood Block Leader
           6 Assistant chief/Group Village Head
           7 Local elder
           8 Traditional Authority (TA)
           9 Government
          10 Neighbor
          11 Politician
          12 Paramount chief
          13 Commercial farmer/investor
          14 Other (specify)
          15 Don't Know/Refuse to answer

.
```

5.1.3 Merging Datasets

Since the LGPI 2019 data is divided into datasets by both topic and region, it is crucial to be able to merge datasets together.

There are two different types of merging you might want to do

1. Different topics for a single region.

   For example, `demographics` and `educ` in Nairobi.

2. Different regions for a single topic.

   For example, `land` in Zambia Border and Malawi Border

Different topics for a single region If you want to merge datasets containing different topics for the same region you need to merge using SbjNum which is the unique household ID. You can do this using the command:

```
use "FILEPATH/dat1.dta"

merge 1:1 SbjNum using "FILEPATH/dat2.dta"
```

where `dat1` is the first dataset and `dat2` is the second dataset. This creates a new dataset called `newdat` is essentially the two original datasets put side-by-side.

NOTE: Put the two datasets in the same directory first.

5.2 R

5.2.1 Importing

- Save the R file (.rds) to your directory of choice.
- Set your working directory to where the file is stored: *setwd("filepathname")*
- Read in the R file: *readRDS("filename.rds")*

5.2.2 Encoding of Variables

If you would like to see all levels of a particular variable use the *levels(x)* function. See the example below.

```
> levels(mydat$LAND_Q8)
 [1] "Village/Town Council"                    "Spouse's parents"
 [3] "Spouse's other relatives"                "Chief"
 [5] "Village Head/Neighborhood Block Leader" "Assistant chief/Group Village Head"
 [7] "Local elder"                             "Traditional Authority (TA)"
 [9] "Government"                              "Neighbor"
[11] "Politician"                              "Paramount chief"
[13] "Commercial farmer/investor"              "Other (specify)"
[15] "Don't Know/Refuse to answer"
```

5.2.3 Merging Datasets

Since the LGPI 2019 data is divided into datasets by both topic and region, it is crucial to be able to merge datasets together.

There are two different types of merging you might want to do

1. Different topics for a single region.

   For example, `demographics` and `educ` in Nairobi.

2. Different regions for a single topic.

   For example, `land` in Zambia Border and Malawi Border

Different topics for a single region

If you want to merge datasets containing different topics for the same region you need to merge using SbjNum which is the unique household ID. You can do this using the command:

```
newdat = merge(dat1,dat2,by = SbjNum)
```

where `dat1` is the first dataset and `dat2` is the second dataset. This creates a new dataset called `newdat` is essentially the two original datasets put side-by-side.

SECTION 6

Data Cleaning

This chapter outlines the procedure used to clean the data from the LGPI 2019.

6.1 Dropped Observations

The reasons that an observation may have been completely eliminated from the dataset are:

1. The observation was incomplete. We defined incomplete observations as those who did not have complete demographic data and/or did not answer both the first and last questions of the survey.

2. The observation, while complete, is located too far from the original sampling frame. We define too far as being more than 1 kilometer away from any square kilometer in the sampling frame.

3. The observation, was not done during the official sampling time period. Kenya/Zambia: May 29, Malawi: Sept 23

4. There was no GPS data available for the observation.

5. The implementation partner told us the observation needed to be dropped due to a confirmed quality issue.

Table 1: Breakdown of Observations

	No GPS Data	Incomplete Observations	Too Far Away	Outside Fielding Period	Partner Drop	Usable Observations
Kenya	2	21	39	1	20	3788
Zambia	2	303	136	10	0	9864
Malawi	2	139	9	69	0	10302

6.2 Don't Know and Refuse to Answer Responses

For many of the questions, both multiple choice and free response, "Don't Know" or "Refuse to Answer" are possible answer choices. Throughout the survey we see this type of answer choice in the following ways:

- Don't Know
- Refuse to Answer
- Don't Know/Refuse to Answer

It should be assumed that if "Don't Know/Refuse to Answer" is given as an answer choice, the designers of the survey determined that there was insufficient insight to recording "Don't Know" and "Refuse to Answer" separately.

7

SECTION

Survey Weights

7.1 Sampling & Post–Stratification Weights

A typical survey has two types of weights, sampling weights and post–stratification weights. We know describe the weights used in the LGPI 2019 survey.

Sampling (also called probability) weights are used to account for the fact that not all individuals in the sampling frame had the same probability of being selected into the sample. Since we utilized a stratified multi-stage probability proportional to size sampling method all individuals had the same probability of being selected and therefore we do not use sampling weights.

Post–stratification weights use auxiliary data to weight sample observations so that distributions of key control variables in the sample reflect the distribution of the same variables in the population. For the LGPI 2019 survey we compute the post–stratification weights separately for each sampling region (Nairobi, Lusaka, Lilongwe, Zambia Border, Malawi Border). We create two version of the weights for each region:

- Version 1: Stratify on Age, Education, Gender
- Version 2: Stratify on Age, Education, Gender, Region*

*–Region == geolevel2 region for each country.

IPUMS International https://international.ipums.org/international/ data was used as the auxiliary data source to compute strata.

7.2 Using the Weights

In order to incorporate the weights into your analysis follow the steps below:

1. Identify what region you are currently analyzing (KE – Nairobi, LU – Lusaka, LI – Lilongwe, MB – Malawi Border, ZB – Zambia Border)
2. Import the starting dataset for the region you want to analyze. Filename of the form: *StartingData_YY.dta* where YY is the region code.
3. Merge additional datasets.
4. Identify whether or not you want to post–stratify on region.
5. Copy the corresponding `svyset` command into your DO-file under your data import/merge statements. The code can be found below and in the *Survey Weights* file.

Lilongwe

```
svyset sqkm , fpc(fpc1) strata(sector) || hect, fpc(fpc2) || _n , singleunit(scaled)
poststrata(STRATA_RAEG_name) postweight(STRATA_RAEG_size)

svyset sqkm , fpc(fpc1) strata(sector) || hect, fpc(fpc2) || _n , singleunit(scaled)
poststrata(STRATA_EAG_name) postweight(STRATA_EAG_size)
```

Malawi Border

```
svyset sqkm , fpc(fpc1) strata(sector) || hect, fpc(fpc2) || _n , singleunit(scaled)
poststrata(STRATA_MB_RAEG_name) postweight(STRATA_MB_RAEG_size)
svyset sqkm , fpc(fpc1) strata(sector) || hect, fpc(fpc2) || _n , singleunit(scaled)
poststrata(STRATA_MB_EAG_name) postweight(STRATA_MB_EAG_size)
```

Lusaka

```
svyset sqkm , fpc(fpc1) strata(sector) || hect, fpc(fpc2) || _n , singleunit(scaled)
poststrata(STRATA_ZB_RAEG_name) postweight(STRATA_ZB_RAEG_size)
svyset sqkm , fpc(fpc1) strata(sector) || hect, fpc(fpc2) || _n , singleunit(scaled)
poststrata(STRATA_ZB_EAG_name) postweight(STRATA_ZB_EAG_size)
```

Zambia Border

```
svyset sqkm , fpc(fpc1) strata(sector) || hect, fpc(fpc2) || _n , singleunit(scaled)
poststrata(STRATA_ZB_RAEG_name) postweight(STRATA_ZB_RAEG_size)
```

```
svyset sqkm , fpc(fpc1) strata(sector) || hect, fpc(fpc2) || _n , singleunit(scaled)
poststrata(STRATA_ZB_EAG_name) postweight(STRATA_ZB_EAG_size)
```

Kenya

```
svyset sqkm , fpc(fpc1) strata(sector) || hect, fpc(fpc2) || _n , singleunit(center)
poststrata(STRATA_RAEG_name) postweight(STRATA_RAEG_size)
svyset sqkm , fpc(fpc1) strata(sector) || hect, fpc(fpc2) || _n , singleunit(center)
poststrata(STRATA_EAG_name) postweight(STRATA_EAG_size)
```

7.3 Checking the Weights

Run the code below to verify that the survey weights are working

```
svydescribe
```

Doing so will produce output like the image below

			#Obs per Unit		
Stratum	#Units	#Obs	min	mean	max
1	15	417	21	27.8	51
2	44	1,167	17	26.5	49
3	20	517	18	25.9	42
4	23	532	1	23.1	32
5	10	246	23	24.6	26
6	22	521	4	23.7	28
7	13	279	4	21.5	26
8	5	88	6	17.6	25
8	152	3,767	1	24.8	51

21 = #Obs with missing values in the survey characteristics

3,788

Figure 1: Output from `svydescribe` command.

When reviewing output like that in Figure 1 we are most concerned with the "#Obs with missing values in the survey characteristics". This number represented the number of observations which will be dropped with using the survey weights. This number should be very small relative to the total number of unweighted observations.

SECTION 8

Sampling Plan of LGPI 2019

The sampling of the LGPI was performed independently in 5 regions across 3 countries: Kenya, Zambia, and Malawi. We name the 5 regions:

- Nairobi
- Lilongwe
- Malawi Border
- Lusaka
- Zambia Border

The maps below show the areas canvased by the LGPI 2019 survey. Figure 2 shows all of the areas included in the LGPI 2019 survey. The bull's eye shaped areas correspond to the Nairobi, Lilongwe, and Lusaka regions. Figure 3 and Figure 4 show the Malawi/Zambia and Kenya sampling areas in more detail.

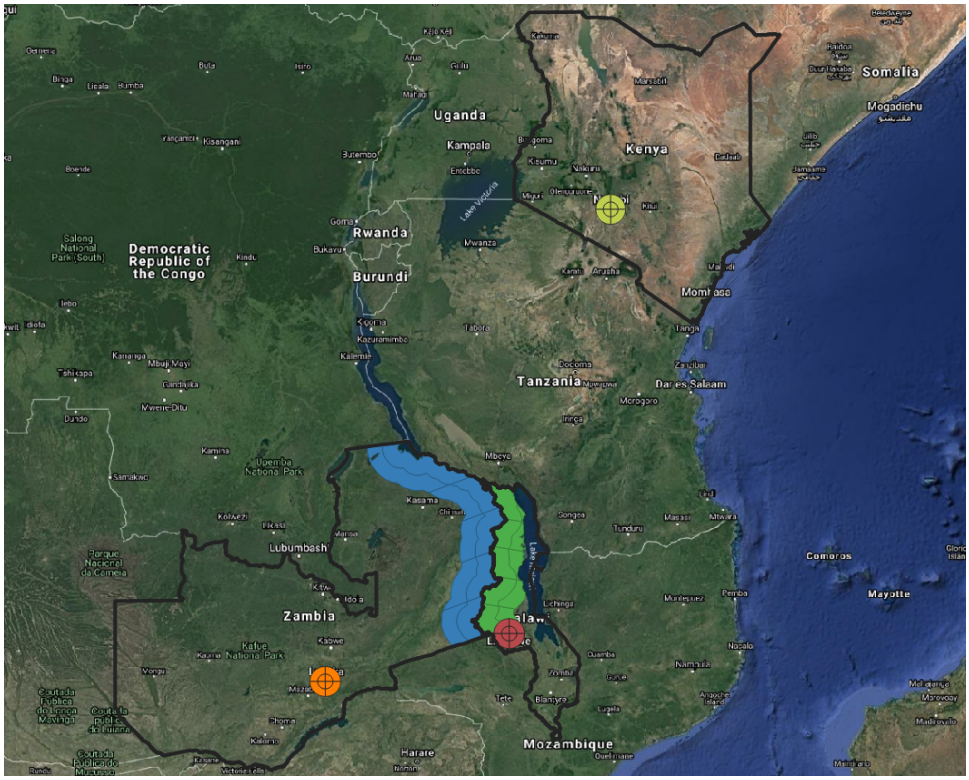


Figure 2: LGPI 2019 Sampling Areas

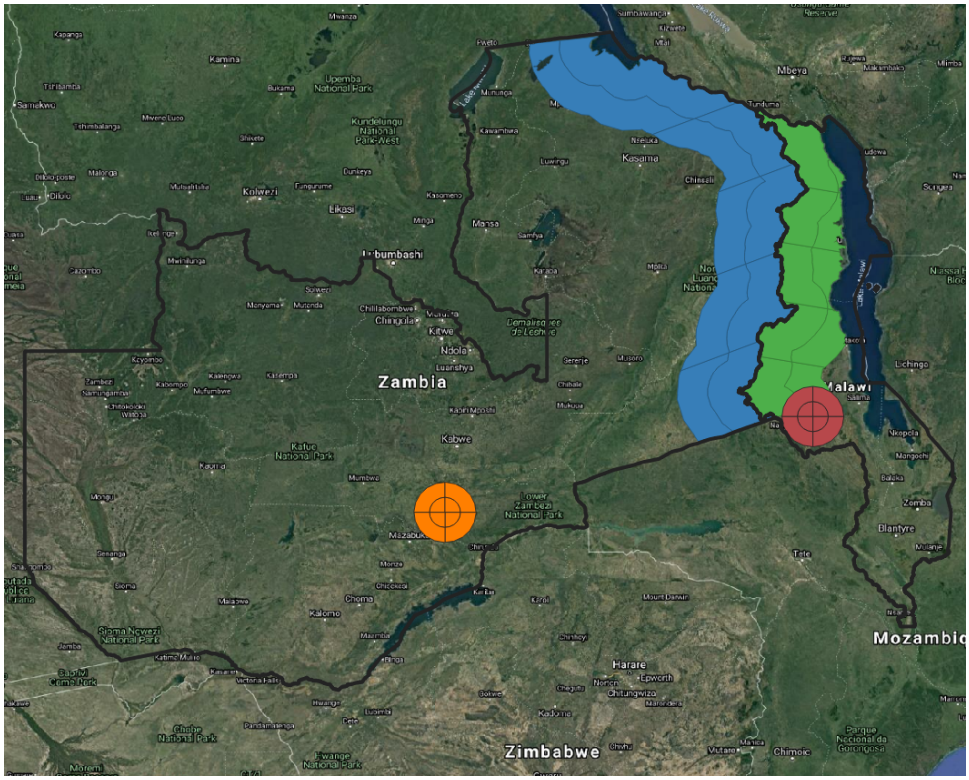


Figure 3: Malawi and Zambia Sampling Areas

Figure 4: Kenya Sampling Areas

As seen in Figure 2, there are two different types of regions in the sample. There are three capital regions (Nairobi, Lilongwe, and Lusaka) and two border regions (Malawi Border and Zambia Border). The LGPI 2019 survey utilized two sampling plans, one for the capital regions and one for the border regions. The two sampling plans are very similar with two major differences:

1. Definition of Bins
2. Selection Method for Choosing Square Kilometers

8.1 Overview of Sampling Plans

The sampling plans for the capital and border regions can be described as a multi-stage probability proportional to size sampling. Each sampling plan follows the same general steps:

1. Define the bins. (these are the strata)
2. Choose square kilometers within the bins using PPS. (stage 1)
3. In each selected sqkm, create a grid of hectares.
4. Remove all empty/underpopulated hectares from consideration.
5. Within each sqkm, randomly assign an order to the viable hectares using PPS (stage 2)

To see all the code used to create the sampling plan visit: https://github.com/senickel/sampling_documentation

and

https://senickel.github.io/sampling_documentation/index.html#prerequisites

The subsequent sections provide a detailed description of the sampling plan.

8.2 Defining the Bins

The LGPI 2019 Survey does not cover the entirety of any of the three countries included in the project. Therefore we must first define where in the country we will be conducting our survey. We call these large defining areas Bins. In the urban areas there are 8 bins and in the rural area there are 10 bins.

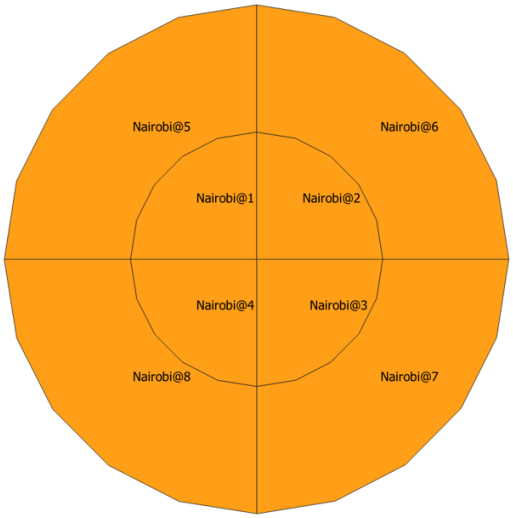
8.2.1 Creating Urban Bins

When surveying in a capital region, we want to ensure sampling of areas close to and further away from the city center. Therefore, we structure the capital region bins as a "bull's eye" over the city center. The bull's eyes are centered at the following locations:

- Nairobi: 36.81667 Longitude, -1.28333 Latitude
- Lilongwe: 33.783333 Longitude, -13.983333 Latitude
- Lusaka: 28.283333 Longitude, -15.416667 Latitude

Each bull's eye consists of two concentric circles: the first is 25km from the center and the second is 50 kilometers from the center. Then a vertical line and a horizontal line are drawn through the city center. The result is a marked off area like Figure 5. As seen in Figure 5, every bin is given a name of the form "RegionName@Bin#".

Figure 5: Sectors for the Nairobi sampling area.



8.2.2 Creating Rural Bins

When defining the bins for the border sampling plans, we employ a similar rule as the one used to create the capital region bins. To ensure that we sample areas both close to and far away from the border, we begin by defining two regions. The first is 50km from the border and the second is 50-100km from the border (as geography allows).

Then to ensure that we have equal sampling in the north-south direction we divide each of the first two regions into 5 smaller regions. The results for the two border regions are shown in Figures 6 and 7.

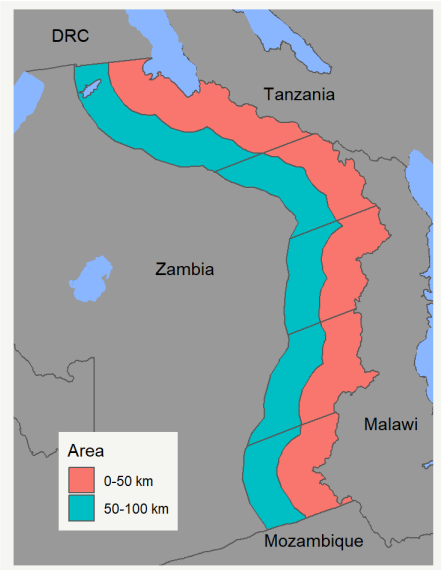


Figure 6: Bins for Zambia Border Region

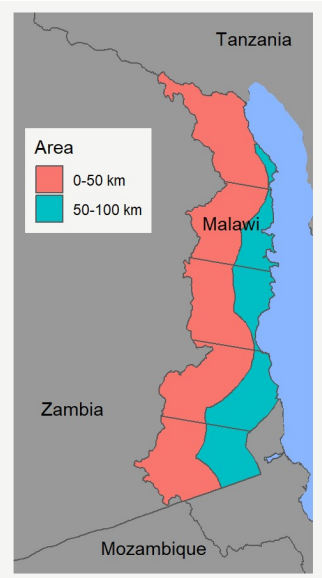


Figure 7: Bins for Malawi Border Region

8.3 Selection of Square Kilometers

The square kilometer sampling unit is used as a proxy of a community. Since boundaries of villages and communities are unclear and in some cases dynamic, we define a square kilometer area to be a community. Sampling these units will allow us to gather data from individuals living in close proximity to each other.

8.3.1 Capital Regions

We perform a stratified sample (stratum = bin) using PPS to select square kilometers (SQKM). First we create a grid of 1 kilometer squares over the bins and compute the expected population density for that areas using WorldPop data. We then sample a predetermined number of squares using probability proportional to size (where size = population density) sampling.

The number of square kilometers sampled was equal to the desired number of communities plus a small number of supplemental units. The supplemental units serve as backups in the case that a chosen square kilometer is not sufficiently populated to obtain the necessary number of observations.

8.3.2 Border Regions

The border regions are notably larger than the capital regions, so traveling through the region will be intensive and expensive for the enumerators. To help mitigate this issue, we we sample 5 kilometer squares instead of sampling individual square kilometers. That is, we sample 25

square kilometers at a time. To do this we create a grid of 5km squares in each bin, and then using PPS select an appropriate number of 5km squares.

8.3.3 Sampled Square Kilometer Evaluation

Once the square kilometers have been sampled, they must be reviewed to ensure they are sufficiently populated. While PPS should minimize the probably of selecting underpopulated areas, there is still a small probability they could end up in the sample. A google map image of each selected square kilometer was reviewed to ensure the area was populated. If the area was not sufficiently populated it was removed from the sample.

8.4 Selection of Hectares

In every verified square kilometer from the previous step, we create a grid of hectares. Each hectares is checked visually to ensure it is populated. We then use PPS to randomly assign an order to the inhabited hectares within each square kilometer.

This second stage of sampling helps ensure that the entire square kilometer is being sampled and thus providing a more complete picture of the community.

8.5 Selection of Houses Respondents

This stage of the sampling plan was implemented on the ground by teams of enumerators. Teams were sent to a specified hectare by their team leader. They were instructed to enter the hectare using tablets to track their locations and confirm they were in the correct area. They would then go to the center of the hectare and then move outward in a random walk. Once a household had been selected (and agreed to participate) a Kish grid was used to randomly choose a respondent from all reported adults (at least 18 years old) in the household.

8.6 Reaching Target Number of Observations

Starting from the center of a hectare, the team of enumerators would go in separate directions to additional houses using a random walk. The team continued surveying in the hectare until at least 8 surveys had been completed. Once the hectare was completed, if the target number of observations for the current square kilometer had not yet been met, the team would move onto the next listed hectare for the current square kilometers as outlined in Subsection 8.4. The teams would continue to complete hectares until the target number of observations for a square kilometer had been met.

The target number of observations for each region were:

	Total # Obs	# of SQKM	# Obs Per SQKM	# Obs Per Hect*
Nairobi	3750	150	25	8
Lusaka	4500	150	30	8
Zambia Border	6000	200	30	8
Lilongwe	4500	150	30	8
Malawi Border	6000	200	30	8

*–One will note that the number of observations per hectare does not equally divide the number of observations per square kilometer. This is due to a last minute change to increase the number of observations per hectare from 5/6 to 8. This was done to increase the chances that enumerators would be able to reach the target number of observations per square kilometer.

8.6.1 Team Size

It is important to note that the size of the teams sent to hectares was large enough to minimize the chances that a small number of enumerators complete all of the observations in a square kilometer. By doing so, we help address the issue of enumerator effects confounding with community (sqkm) effects.

8.6.2 Special Considerations

We now describe additional sampling rules that were implemented to help mitigate issues in the field.

One–Hectare–West

Justification: We added this rule as a way to combat the issue of low population density areas.
Rule: In the event that the team of enumerators could not obtain 8 observations in a given hectare, they were allowed to move one hectare west to make up the missing observations.
Implementation Period: This rule was in place in all regions throughout the entirety of the respective fielding period.

Adjacent–Square–Kilometer

Justification: During fielding in Kenya and Zambia it become apparent there were some issues with the population density of selected and rejected areas. Namely that many selected areas

were not sufficiently populated and numerous rejected areas were in fact populated. This resulted in a large number of incomplete (number of obs < 30/25) square kilometers.

Rule: If after an entire square kilometer has been exhausted, the square kilometer is still not complete, then enumerators may go to any *adjacent* sqkm not in the sampling plan to obtain the remaining observations. Enumerators are instructed to obtain the observations as close to the border as possible.

Implementation Period: This rule was first implemented in Kenya and Zambia on July 27, 2019. It was implemented during the entire Malawi fielding.

Open–Square–Kilometer

Justification: During fielding in Kenya and Zambia we learned that many of the selected hectares were underpopulated while some of the rejected hectares were populated.

Rule: If after all hectares in a given sqkm have been exhausted the target number of observations for that sqkm have not been met, enumerators may go anywhere within the sqkm

Implementation Period: This rule was never used in Kenya. It was first implemented in Zambia on September 5, 2019. It was implemented during the entire Malawi fielding.