Sample Design and Weighting Procedures for Rural Mon State Household Survey

David J. Megill
Sampling Consultant, Michigan State University
May 2015

1. Background and Objectives

The main objective of the Rural Mon State Household Survey was to study the agricultural and socioeconomic characteristics of the rural households of the Mon State, with a focus on the following activities: rubber, rice, orchards and marine fishing. The landless households were also included in the survey since they are a subgroup of special interest to the project. Therefore most of the rural households in the Mon State are in scope for this survey.

The sampling frame for the Rural Mon State Household Survey was based on the preliminary data and maps from the 2014 Population and Housing Census of Myanmar. Auxiliary information was used to classify the village tracts in Mon by the level of intensity of the different crop and marine fishing activities of interest.

The purpose of this report is to describe the sample design and weighting procedures for the Rural Mon State Household Survey. The sampling methodology was developed in collaboration with Ellen Payongayong, Michigan State University Consultant. The sampling frame was developed with the collaboration of Yin Yin Kyaing, Deputy Director, Department of Population, Ministry of Immigration and Population.

2. Sampling Frame and Stratification for Rural Mon State Household Survey

A stratified two-stage sample design was used for the Rural Mon State Household Survey. The sampling frame was based on preliminary summary data and maps from the 2014 Population and Housing Census of Myanmar. Based on the survey objectives, the sampling frame was limited to the rural households in the Mon State. The primary sampling units (PSUs) selected at the first sampling stage were the census enumeration areas (EAs), which are segments defined within the village tracts and wards for the purposes of the data collection for the 2014 Census. The EAs which only had institutional population were excluded from the frame. The rural EAs in the Mon State have an average of 132 households, which is a practical size for the purpose of conducting a listing operation to update the frame of households in the sample EAs. The original frame from the 2014 Myanmar Census included 2,256 rural EAs for the Mon State.

In order to examine the distribution of the sampling frame of EAs by predominant crop or fishing activity, each village tract in the Mon State was classified into three categories for each crop or marine fishing activity that was present: (1) low, (2) medium and (3) high. This information on the activities and crop levels was merged with the sampling frame of EAs from the 2014 Myanmar Census, which included the preliminary number of households in each EA.

The information on the crops and marine fishing activities was used to stratify the sampling frame of rural EAs for the Mon State in order to improve the efficiency of the sample design and ensure an effective allocation of the sample to cover the different crop and marine fishing activities. Since most of the EAs in the frame had more than one activity at different levels, it was first necessary to identify the predominant crop or marine fishing activity for each EA. This required an iterative approach of tabulating the distribution of EAs by crop and marine fishing activities in different ways. Table 1 shows the distribution of the Mon rural sample EAs by activity and level categories.

It should be noted that the classification by crop activity was carried out at the village tract level, so an EA within the village tract may or may not have the particular activity being classified. However, it is expected that this classification should be effective overall for the purposes of stratification.

Table 1. Distribution of the Mon rural EAs in the sampling frame by crop/fishing activity and level

	Number of EAs by level				% of EAs in
Activity	1-Low	2-Medium	3-High	Total	frame
Rubber	619	575	639	1,833	81.3%
Rice	329	655	1,166	2,150	95.3%
Orchards	570	351	338	1,259	55.8%
Marine fishing	351	246	109	706	31.3%

It can be seen in Table 1 that the overall most predominant crop is rice, which is found in 2,150 of the original 2,256 EAs in the frame (or 95.3%), and more than half of those EAs are in the high level category. The least frequent activity is marine fishing, which is only found in 31.3% of the EAs, and only 15.4% of those EAs are in the high level category.

Next we examined the distribution of rural EAs in the frame by the different combinations of crops and fishing activities. This is shown below in Table 2. It can be seen in Table 2 that most of the rural EAs in the Mon sampling frame have multiple activities. After excluding the 9 EAs which did not have any of the activities of interest and a few EAs that did not have information on the number of households in the preliminary Census data, the final sampling frame of rural EAs used for the Rural Mon State Household Survey had a total of 2,235 EAs.

Table 2. Distribution of the Mon rural EAs in the sampling frame by combinations of crop and marine fishing activities

Activities	No. of EAS
None (i.e., freshwater area)	9
Orchards and marine fishing	33
Rubber and orchards	64
Rice only	197
Rice and orchards	74
Rice and marine fishing	47
Rice, orchards and marine fishing	63
Rice and rubber	588
Rice, rubber and orchards	618
Rice, rubber and marine fishing	156
Rice, rubber, orchards and marine fishing	407
Total rural EAs	2,256

Based on the distribution of the frame of rural EAs for the Mon State shown in Tables 1 and 2, hierarchical criteria were used to define one stratum for each predominant activity. The stratification criteria were designed to give priority to the less frequent activities in order to ensure a sufficient number of observations for each activity in the final sample of households selected for the survey. The stratification criteria are specified in Table 3.

Table 3. Hierarchical criteria used for defining stratum for each predominant activity

Stratum	Activity	Criteria
1	Marine fishing	All marine fishing in level categories 2 and 3
2	Orchards	Orchards in level categories 2 and 3, and not in Stratum 1
3	Rubber	Rubber category >= rice category, and not in Strata 1 or 2
4	Rice	Rice category > rubber category, and not in Strata 1 or 2

In order to improve the effectiveness of the stratification, each stratum was further divided into two categories (low and high) for the corresponding activity. For this purpose the original level categories 1 and 2 were combined to form the low substratum for each predominant activity, and the high substratum corresponded to the original level category 3. Following the coding of the strata based on these specifications, the final distribution of the frame of rural EAs for the Rural Mon State Household Survey is shown in Table 4.

Table 4. Final distribution of sampling frame of rural EAs in the Mon State by predominant activity stratum and level substratum

Stratum	Predominant activity	Substrata		Total
		1-Low	2-High	
1	Marine fishing	244	109	353
2	Orchards	227	241	468
3	Rubber	93	398	491
4	Rice	104	819	923
Total		668	1,567	2,235

3. Sample Size and Allocation for Rural Mon State Household Survey

The sample size for the Rural Mon State Household Survey depends on the survey objectives as well as the overall budget constraints. One of the objectives was to ensure a sufficient level of precision for the indicators of each of the crop and marine fishing activities. The overall sample size could not exceed 2,000 households, so it was important to allocate the sample strategically to the different strata to ensure a sufficient number of sample households for each of the activities covered by the survey.

It is also important to determine the most effective number of sample households to select per cluster (EA). In order to ensure that the design effects from the clustered sample are not too high, the optimum number of sample households per cluster for this type of socioeconomic survey is generally within the range of 8 to 15 households. Taking into consideration that there will be a certain level of nonresponse, it was decided to select 12 households per sample EA. A sample of 35 sample EAs per stratum will ensure a reasonable dispersion of the sample within each stratum. Within each activity stratum the 35 sample EAs were allocated to the level substrata in proportion to the distribution of the frame, while giving a weight of 2 to the high level stratum. This will increase the sampling rate for the high level strata to improve the coverage of each activity.

The final sample size was 140 EAs and 1,680 households, with 420 sample households per predominant activity stratum. This sample size should provide a reasonable level of precision for the indicators by activity, especially since many sample households will be involved in more than one activity. For example, rice is found in all the sampling strata. The final distribution of the sample EAs and households by stratum and substratum is shown in Table 5.

Table 5. Allocation of sample EAs and households for Rural Mon State Household Survey by activity stratum and level substratum

Stratum	Predominant activity	Total		1-Low substratum		2-High substratum	
		Sample EAs	Sample households	Sample EAs	Sample households	Sample EAs	Sample households
1	Marine fishing	35	420	18	216	17	204
2	Orchards	35	420	12	144	23	276
3	Rubber	35	420	4	48	31	372
4	Rice	35	420	3	36	32	384
Total		140	1,680	37	444	103	1,236

4. Sample Selection Procedures for Rural Mon State Household Survey

Within each predominant activity stratum and level substratum the number of sample EAs specified in Table 5 was selected systematically with probability proportional to size (PPS), where the measure of size was based on the number of households in the 2014 Census frame. First the list of EAs within each substratum were ordered geographically to provide additional implicit stratification. The systematic selection of the EAs with PPS within each substratum involved the following steps:

1) Cumulate the measures of size (number of households) down the ordered list of EAs within the substratum. The final cumulated measure of size is the total number of households in the frame for the substratum (M_h).

2) To obtain the sampling interval for substratum h (I_h), divide M_h by the total number of EAs to be selected in substratum h (n_h), specified in Table 5.

3) Select a random number (R_h) between 0.01 and I_h. The sample EAs in substratum h will be identified by the following selection numbers:

   $S_{hi} = R_h + [I_h \times (i - 1)]$, rounded up,

   where i = 1, 2, ..., n_h

   The i-th selected PSU is the one with the first cumulated measure of size that is greater than or equal to S_{hi}.

The selection of PSUs using systematic PPS sampling by substratum was implemented using the Complex Samples module of SPSS. For the SPSS Complex Samples application it is necessary to specify the sample selection method (systematic PPS); the variables for the stratum, PSU and measure of size; and the number of PSUs to be selected in each stratum. The stratum code was defined as the concatenation of the predominant activity code (1 to 4) and the level substratum

code (1 to 2). The SPSS software generates a new database with a record for each selected EA that includes the sampling frame information and the first stage weight for each sample EA.

A new listing of households will be conducted in each sample EA. If some households are considered out of scope, they can be screened out at the listing stage once the corresponding questions are included in the listing sheet. Since the landless households are considered eligible for selection, almost all of the households will be included in the frame. A random systematic sample of 12 households will be selected from the listing for each sample EA. This second stage selection procedure will involve the following steps:

1) All the households in valid (occupied) housing units should be assigned a serial number from 1 to M'_{hi}, the total number of households listed in the EA.

2) To obtain the sampling interval for the selection of households within the sample EA (I_{hi}), divide M'_{hi} by 12, and maintain 2 decimal places.

3) Select a random number (R_{hi}) with 2 decimal places, between 0.01 and I_{hi}. The sample households within the sample EA will be identified by the following selection numbers:

   $S_{hij} = R_{hi} + \left[I_{hi} \times (j-1)\right]$, rounded up,

   where j = 1, 2, 3,..., 12

The j-th selected household is the one with a serial number equal to S_{hij}.

An Excel spreadsheet was developed for generating this random systematic selection of households, with formulas for calculating the interval, generating the random start and calculating the selection numbers. This spreadsheet was used to produce a household selection table that was used in the field to identify the serial number of the in-scope sample households in reference to the total number of households listed in each sample EA.

This second stage household selection procedure will provide a representative sample of all households within each sample EA, which will cover the predominant and secondary activities conducted by these households. Based on the stratification of the sample EAs by predominant activity, this should ensure a sufficient number of sample households for the predominant activities. As mentioned previously, many households are involved in more than one activity.

5. Weighting Procedures for Rural Mon State Household Survey

In order for the sample estimates from the Rural Mon State Household Survey to be representative of the population, it is necessary to multiply the data by a sampling weight, or expansion factor. The basic weight for each sample household would be equal to the inverse of its probability of selection (calculated by multiplying the probabilities at each sampling stage).

The sampling probabilities at each stage of selection can be maintained in an Excel spreadsheet with the information from the frame for the sample EAs within each substratum.

A stratified two-stage sample design was used for the Rural Mon State Household Survey. The overall probability of selection for sample households can be expressed as follows:

$$p_{hi} = \frac{n_h \times M_{hi}}{M_h} \times \frac{m_{hi}}{M'_{hi}},$$

where:

p_{hi} = probability of selection for the sample households in the i-th sample EA of substratum h

n_h = number of sample EAs selected in substratum h

M_h = total number of households in substratum h

M_{hi} = number of households in the frame for the i-th sample EA of substratum h

m_{hi} = 12 = number of sample households selected in the i-th sample EA of substratum h

M'_{hi} = total number of eligible households in the updated listing for the i-th sample EA of substratum h

The two components of this probability of selection correspond to the individual sampling stages. The basic sampling weight for the sample households is calculated as the inverse of this probability of selection. Based on the previous expression for the probability, the weight for the sample households can be simplified as follows:

$$W_{hi} = \frac{M_h \times M'_{hi}}{n_h \times M_{hi} \times m_{hi}},$$

where:

W_{hi} = basic weight for the sample households in the i-th sample EA of substratum h

These weights will vary based on the difference between the number of eligible households in the updated listing for each sample EA and the corresponding number of households in the 2014 Census frame. Following the data collection for the Rural Mon State Household Survey, it will be necessary to adjust these basic weights to take into account any non-interviews. The weight for the sample households in each sample EA would be adjusted as follows:

$$W'_{hi} = \frac{M_h \times M'_{hi}}{n_h \times M_{hi} \times m_{hi}} \times \frac{m_{hi}}{m'_{hi}} = \frac{M_h \times M'_{hi}}{n_h \times M_{hi} \times m'_{hi}}$$

where:

W'_{hi} = adjusted weight for the sample households in the i-th sample EA of substratum h

m'_{hi} = number of sample households with completed interviews in the i-th sample EA of substratum h