Proposed sample design and weighting procedure for the Liberia LFS/CWIQ 2009

Peter Wingfield-Digby, ILO statistical consultant, 24 June 2009

Introduction

The Liberia Institute of Statistics and Geo-Information Services (LISGIS) plans to conduct a national household survey towards the end of 2009, which will incorporate both a Labour Force Survey (LFS) and a Core Welfare Indicators Questionnaire (CWIQ). The design of the sample can be based on the national sampling frame developed from the recently conducted 2008 Population Census, since the detailed results of the census are now being made available.

Domains of analysis

With the resources available for the two surveys being combined into one survey, it should be possible to select a larger sample than would have been possible if the two surveys had been done separately. This sample needs to allow for the presentation of estimates at the county level, which are required for assessing progress under the Poverty Reduction Strategy (PRS). The earlier CWIQ and Demographic and Health Survey (DHS), both conducted in 2007, had smaller samples which only allowed for estimates at the regional level. Artificial regional groupings were constructed as follows:

North Central	- Bong, Nimba, Lofa
North Western	- Bomi, Grand Cape Mount, Gbarpolu
South Central	- Montserrado (outside Monrovia), Margibi, Grand Bassa
South Eastern A	- River Cess, Sinoe, Grand Gedeh
South Eastern B	- River Gee, Grand Kru, Maryland
Greater Monrovia	

Separate urban and rural estimates were produced from the data collected in those two surveys. In the case of the DHS, those estimates were provided only at the national level. In the case of the CWIQ, an attempt appears to have been made to provide separate estimates for urban and rural households in each region (although the size of the urban samples - barely more than 100 households - does not in fact warrant this level of detail).

CWIQ 2007 used exactly the same sample design as that used for DHS 2007. A total of 300 sampling points (enumeration areas - EAs) were selected around the country, of which 116 were urban and 184 were rural. The only major difference was that for the CWIQ only 12 households were selected in each EA, whereas for the DHS 25 households had been selected.

While the sample for the 2009 LFS/CWIQ has been designed to allow for separate estimates for each county, it is not possible to have separate urban and rural estimates at the county level as well. To do so would require a much larger sample, and in any case there are several counties (notably River Cess, Grand Kru and Gbarpolu) which have very few urban EAs, so that it is difficult to get adequate urban samples. The urban/rural breakdown will therefore only be possible at the regional level.

CWIQ / LFS sampling considerations

With the preliminary results of the 2008 population census now available, it is possible to see the marked contrast between the sampling frame used for the earlier DHS and CWIQ surveys and the frame that has become available for the LFS/CWIQ 2009. These differences are highlighted in Table 1, which shows the number of urban and rural EAs in each county. The major differences of interest to the sample designer are as follows.

While the overall number of EAs has increased by just over 50 percent, there is a marked contrast in what has happened to urban and rural EAs: the number of urban EAs has almost trebled, while the number of rural EAs has increased by less than 10 percent. This increase in urban EAs is due mainly to two factors: first, the movement of people from rural to urban areas; and secondly, a change in the designation of what counts as an urban area. In the 1984 Census, urban areas in each county “consisted mainly of the county capitals”. For the 2008 Census, a much broader definition was used, with all settlements of 2000 or more persons being counted as urban. The ratio of urban to rural EAs has thus changed from 25:75 in 1984 to 46:54 in 2008.

It is also clear from the table that the trends have not been consistent across all counties. Only one county (Sinoe) saw a drop in the number of urban EAs it contains, whereas several other counties saw massive increases in the number of urban EAs.

Table 1: Number of urban and rural enumeration areas by county, 1984 and 2008 Censuses

	Number of urban and rural EAs						Percentage change		
	1984 Census			2008 Census			1984-2008		
County	***Urban***	***Rural***	***Total***	***Urban***	***Rural***	***Total***	***Urban***	***Rural***	***Total***
Bomi	43	146	189	54	214	268	26	47	42
Grand Cape Mount	11	160	171	22	246	268	100	54	57
Gbarpolu	2	137	139	15	127	142	650	-7	2
Montserrado (ex GM)	6	165	171	100	180	280	1,567	9	64
Margibi	50	229	279	144	282	426	188	23	53
Grand Bassa	85	324	409	128	336	464	51	4	13
River Cess	4	90	94	5	143	148	25	59	57
Sinoe	38	96	134	23	193	216	-39	101	61
Grand Gedeh	35	145	180	73	100	173	109	-31	-4
Rivergee	1	108	109	27	81	108	2,600	-25	-1
Grand Kru	7	93	100	9	119	128	29	28	28
Maryland	28	119	147	64	107	171	129	-10	16
Bong	58	482	540	253	667	920	336	38	70
Nimba	28	775	803	171	600	771	511	-23	-4
Lofa	51	378	429	134	365	499	163	-3	16
Greater Monrovia	708	0	708	1,952	0	1,952	176	-	176
Total	**1,155**	**3,447**	**4,602**	**3,174**	**3,760**	**6,934**	175	9	51

Table 2 shows the distribution of urban and rural households by county and by region, according to the Preliminary Results of the 2008 Census.

Table 2: Number of urban and rural households by county and region, 2008 Census

County	Urban	Rural	Total	Region	Urban	Rural	Total
Bomi	4,172	16,538	20,710				
Grand Cape Mount	1,940	22,267	24,207	North Western	7,776	51,940	59,716
Gbarpolu	1,664	13,135	14,799				
Montserrado (ex GM)	12,946	18,906	31,852				
Margibi	18,024	27,513	45,537	South Central	43,389	82,124	125,513
Grand Bassa	12,419	35,705	48,124				
River Cess	511	13,679	14,190				
Sinoe	2,609	13,362	15,971	South Eastern A	10,089	38,451	48,540
Grand Gedeh	6,969	11,410	18,379				
River Gee	2,574	7,566	10,140				
Grand Kru	609	8,393	9,002	South Eastern B	10,865	27,749	38,614
Maryland	7,682	11,790	19,472				
Bong	20,976	49,583	70,559				
Nimba	19,675	62,077	81,752	North Central	55,362	146,939	202,301
Lofa	14,711	35,279	49,990				
Greater Monrovia	202,036	-	202,036	Greater Monrovia	202,036	-	202,036
Total	**329,517**	**347,203**	**676,720**		**329,517**	**347,203**	**676,720**

The initial overall assumption is that for the 2009 CWIQ, household estimates will be required at the county level, but separate urban and rural figures will only be given at the regional level. For simplicity, let us assume that in each county 16 urban and 16 rural EAs are selected. For Greater Monrovia an assumed 60 EAs is used. Let us also assume that 12 households are selected per EA and that a team consists of four interviewers and a supervisor and driver, as was the case in the 2007 CWIQ.

Given this scenario, the 32 EAs in each county would yield 384 households. The separate urban and rural estimates would each be based on 48 EAs in each region, giving 576 households. If Monrovia is treated as a separate region and allocated 60 EAs, the total EAs to be covered in the survey would be 540, giving a total of 6480 households. Although these estimates are in terms of households, which is the unit of analysis for much of the data in the CWIQ, a sample of this size would provide an excellent sample for the LFS, which concentrates on the individual rather than the household as its unit of analysis.

Two key factors to take into account are the total length of the fieldwork period and the length of time it takes a team to cover one EA. A reasonable assumption would be that a team could complete an EA in three days, where this time covers not only the work of interviewing and checking questionnaires, but also the time taken to get from one EA to the next, and the time required for listing of all households in the EA. If a team is to cover 16 EAs in all, then the fieldwork would last almost seven weeks (3 x 16 days), assuming that the teams work full-time, seven days a week.

With this arrangement, two teams would be needed to cover each county, and four teams would be needed for Monrovia. These 34 teams would contain a total of 136 interviewers and 34 supervisors. In practice, it would probably be best to distribute the urban and rural EAs in each county between the two teams, rather than asking one team to do all the urban EAs and the other to do all the rural ones.

The proposed sample design

Table 3 shows the proposed distribution of sample EAs by county. (The sample design used for the 2007 DHS and the 2007 CWIQ is also shown, for purposes of comparison.) It has not always been possible to have exactly 16 urban and 16 rural EAs in a county, because of the problem with some counties containing very few urban areas. This is a particular problem for Gbarpolu, River Cess and Grand Kru. In these cases the small number of urban EAs included in the survey is counterbalanced by taking a larger number of rural EAs, but the total number of EAs is still lower than the expected value (32). This total has been allowed to vary from 28 up to 36 EAs. This means that the total sample size in a county varies between 336 and 432 households. The sample size in Monrovia is 720 households. The number of urban and rural EAs at the regional level - which is the domain for which urban and rural estimates will be given - has been maintained at the level of 48 EAs (576 households)

A surprising feature of this design is that (by coincidence) the overall proportion of EAs selected in 2009 will be almost identical to the proportion selected in 2007. For DHS and CWIQ 2007, 10 percent of urban EAs and 5 percent of rural EAs were selected. For LFS/CWIQ 2007, the corresponding percentages are 9 percent and 6 percent. Despite these apparent similarities, the actual number of households to be interviewed will increase from 3600 to 6480, an 80 percent increase.

Selection of EAs in each county

LISGIS maintains a complete listing of all EAs by county. This list has now been split into two parts for each county: urban and rural. Greater Monrovia, which is to be treated as a separate stratum in the sample design, should be separated out from the rest of Montserrado county. Selection will be done separately from among the urban and rural listings for each county. At the first stage, the required number of EAs will be selected in each stratum with probability proportional to size (PPS), where the measure of size is the number of households listed in the census. At the second stage, a fixed number of households (12) will be taken by systematic sampling within each EA that was picked at the first stage. The advantages of using this PPS approach are three-fold:

a) it will result in samples within each stratum that are approximately self-weighting;
b) it helps us to know in advance what the total size of the sample will be; and
c) it provides equal sized workloads within each EA, which makes it easier to organize fieldwork.

Table 3: Samples selected for DHS and CWIQ 2007, and proposed for LFS/CWIQ 2009

	DHS and CWIQ 2007			***Proposed design for LFS/CWIQ 2009***					
	No. of EAs selected			No. of EAs selected					
County	***Urban***	***Rural***	***Total***	***Urban***	***Rural***	***Total***		***Urban***	***Rural***
Bomi	4	4	8	24	12	36			
Grand Cape Mount	2	8	10	16	16	32	North western	48	48
Gbarpolu	2	26	28	8	20	28			
Montserrado (ex GM)	2	10	12	16	16	32			
Margibi	4	15	19	16	16	32	South central	48	48
Grand Bassa	4	11	15	16	16	32			
River Cess	2	6	8	3	25	28			
Sinoe	4	12	16	16	16	32	South eastern A	48	48
Grand Gedeh	7	15	22	29	7	36			
Rivergee	1	11	12	16	16	32			
Grand Kru	2	5	7	6	22	28	South eastern B	48	48
Maryland	8	19	27	26	10	36			
Bong	3	12	15	16	16	32			
Nimba	2	20	22	16	16	32	North central	48	48
Lofa	3	10	13	16	16	32			
Greater Monrovia	66	0	66	60	0	60	Greater Monrovia	60	0
Total	**116**	**184**	**300**	**300**	**240**	**540**	**Total**	300	240

Margibi Urban was used as an illustration, to demonstrate to LISGIS and MoL staff how this would be done. The professional staff of LISGIS are already fairly familiar with this procedure, from having done sampling work before and from having attended a recent workshop on sampling. In the 2008

Census listing, Margibi is shown as containing 144 urban EAs, with a total of 18,024 households. The EAs are placed in order according to their EA number, which provides an implicit geographic stratification within counties, because of the way that EA numbers are assigned. First, the households in the EAs are cumulated; since the EA sizes are 148, 132, 80, 213, etc., the cumulated values are 148, 280, 360, 573, etc. ending with 18,024, the correct sum of all the EA sizes. These cumulated values are placed in a column, with each value alongside its corresponding EA.

From Table 3, it can be seen that we require 16 EAs in Margibi Urban. The appropriate sampling interval is therefore 18024/16=1126.5. A random start is then picked between 0 and 1126.5. The preferred method is to use the function RANDBETWEEN in Excel to find the random start directly.[1] Suppose it comes to 754 (the RANDBETWEEN function only allows for the selection of integers). This number then identifies the EA ending with 0311, which has a cumulated value of 916. The previous EA could not have been picked, since its cumulative value was 673 which is less than 754. The serial number of the selected EA and its corresponding measure of size should be carefully recorded. The sampling interval is then added to this value of 754, and this will identify the second EA to be selected. This process is continued until all 16 EAs in Margibi Urban have been selected. A useful check that no errors have been made is to add the sampling interval to the 16th selection value. If the total number of households (18,024) is subtracted from this value, we should arrive back at the first selection point, because of the circular nature of the selection process.

A similar selection process should be carried out in each stratum, but using the appropriate total size of that stratum (in terms of its total households) and the required number of EAs to be selected (as shown in Table 3).

One issue that might arise in a few cases is that a large EA might be picked twice or more times in the selection process. For instance, if an EA in Margibi Urban has more than 1126 households (the sampling interval), it is bound to be picked once and could be picked twice. However, the EAs have been created only recently for census purposes, and are designed to provide a satisfactory workload for census enumerators (ideally, 80 to 120 households). It is therefore very unlikely that there will be EAs in Margibi Urban with as many as 1126 households.

Calculation of appropriate weights for each EA

While the sample is approximately self-weighting within each stratum, it is not so across the strata. It is therefore necessary to calculate appropriate weights, which can be applied to each household during the data processing, before tables are created. These weights need to take account of the selection probabilities at the two stages. The weight should also take account of the effects of two other things that happen in the field: any differences in the size of the EA, as revealed through the listing operation that will be done before the households are selected for interview, and any non-response (e.g. refusals or non-contacts) that occurs on the survey.

The calculation of this weighting factor can be illustrated for Margibi Urban. The sample design allows for the selection of 16 EAs in Margibi Urban, which will provide a sample of 16 x 12 = 192 households. The total number of census households found in Margibi Urban was 18,024. Let us assume that the first selected EA has 148 households, but that during the listing operation 160 households are found. Let us also assume that for some reason the interviewer fails to interview one of the 12 selected households, and that no replacements are allowed.

The selection probability at the first stage is therefore: 16 x 148 / 18024
The selection probability at the second stage is: 12 / 160
The proportion of households responding successfully is: 11 / 12

The weighting factor is then calculated by multiplying together the inverses of these ratios. This is:

[1] Alternatively, if RAND is used to obtain a random number between 0 and 1 (as was done in the test run), the value found must be multiplied by the sampling interval to get the value of the random start. It is a good idea to write this random number down on a separate piece of paper, to avoid the risk of forgetting it (since the value on the screen tends to be volatile). In the example, the random start came to 0.669947, which multiplied by 1126.5 gave 754.7. This gives us our first selection, which is the EA ending in 0311, since its cumulated value is 916 while the cumulated value of the previous EA was too small (673).

{18024/(16x148)}x(160/12)x(12/11) which can be rewritten as: {18024/(16x12)}x(160/148)x(12/11). The first ratio represents the grossing-up factor required to get from the planned sample in that stratum (192) to the total number of census households in that stratum (18024). The second ratio represents the adjustment required because of differences between the census size measure used in the original sample selection (148) and that found during the later listing exercise (160). The third ratio represents the adjustment required for non-response.

A spreadsheet should be prepared for the calculation of these weights. Down the left will be the names of the various strata for this survey: Bomi urban, Bomi rural, Grand Cape Mount urban, Grand Cape Mount rural, etc. The second column gives a listing of the 540 EAs selected for the survey, according to the stratum in which they are located. The other columns of the spreadsheet will be:

3. Stratum households (e.g. showing 18,024 in the case of Margibi Urban)
4. Stratum sample size (e.g. 192 in the case of Margibi Urban)
5. Weight 1 (col.3 divided by col.4). Note that all selected EAs in a stratum will have the same values in col. 3 and col.4, so the value of Weight 1 will be the same
6. EA size (according to the listing operation)
7. EA size (from the 2008 Census)
8. Weight 2 (col.6 divided by col.7).
9. Planned ‘take’ within each selected EA - this will always be 12 households.
10. Actual ‘take’ within each selected EA - this will be 12 or some lower number.
11. Weight 3 (col.9 divided by col.10).
12. Overall weight = weight 1 x weight 2 x weight 3

As can be seen from the above description, it is essential to keep good records of all the items needed for calculating these weights. In particular, a record must be kept of the two size measures obtained for each EA, the first one coming during the selection of the EAs for the survey, and the second coming from the listing operation in the listing operation in the field.
