Appendix I

Mongolia Socioeconomic Survey – 2007
Sampling design and implementation

Introduction

This appendix describes the sampling strategy adopted by the 2007 Mongolia Socioeconomic Survey (SES) and the details of its practical implementation until immediately before the survey was fielded, on July 1st 2007. It supersedes another two documents on sampling prepared in the course of the past six months[1], reproducing the parts of them that do not deserve amendments as a result of recent circumstances.

Outline of the design

Sample sizes and strata

A total sample of 11,232 households was allocated into three major strata as follows:

- Ulaanbaatar .....3,600 households
- Aimag Centers .....2,640 households
- Rural Area.....4,992 households

The sample was implicitly allocated by districts and horoos[2] in Ulaanbaatar, and by aimags in the Rural Area. Each aimag center was an explicit sub-stratum, with 240 households allocated to Darhan-Uul and Orhon, and 120 households to each of the other aimag centers. The Govisumber aimag was an explicitly excluded stratum.

Sampling stages

The selection strategy was different in each of the three major strata:

In Ulaanbaatar the sample was selected in two stages:

- First, 360 khesegs.
- Second, 10 households in each kheseg.

---

[1] Muñoz, J. *Sampling design of the Mongolia Integrated Household Income and Expenditure Survey and Living Standards Measurement Survey*. Mission reports (December 2006 and April 2007.)

[2] Mongolia is divided into 22 *aimags*. The largest of them – Ulaanbaatar – is subdivided into 9 districts, 121 *horoos* and 1,035 *khesegs*. Each kheseg has approximately 200 households. The rest of the country is divided into *soums* and *bags*. One of the soums in each aimag is normatively considered as the *Aimag Center* and the others as the *Rural Area*.

In The Aimag Capitals, in two stages:

- First, 12 or 24 bags in each aimag center.
- Second, 10 households in each bag.

In the Rural Areas, in three stages:

- First, 52 soums.[3]
- Second, 12 bags in each soum.
- Third, 8 households in each bag.

In the first sampling stage, area units (khesegs, bags or soums) were selected with probability proportional to size (PPS,) using as a measure of size the number of households at the time of the 2005 administrative registration.

Subsequent sampling stages used as sample frame updated lists of households compiled and computerized by the NSO aimag or district offices. In rural soums, administrative records generally sort the *suurin* (sedentary population) and the *malchid* (nomadic herders,) into separate bags, but even when that was not the case (that is, when a bag contained both *suurin* and *malchid*,) each category was still considered separately for sampling purposes (in other words, each mixed bag was conceptually split into a suurin sub-bag and a malchid sub-bag.) This was also done in the few aimag centers that also reported the presence of herders.

The household lists contain the address, size, and name of the head of each household, and whether any of its members is known to work as a non-agricultural self-employed. The group of 8 (or 10) households to be visited by the survey in each kheseg or bag (hereafter referred to as a *cluster*) was selected from these lists as follows:

- In non-herder units, by systematic, unequal probability sampling, giving the households assumed to contain self-employed twice as many chances of being selected than the rest of the households in the unit.

- In herder bags, by equal probability, circular cluster sampling.

Reserve households

In addition to the 8 (or 10) households targeted by the survey, three extra households were selected for each cluster, with the intention of being used as a reserve for eventual nonresponse among the target households. This was done in practice by selecting 11 (or 13)-household clusters first, and then three among them by systematic equal probability sampling to make the reserve.

Multiple clusters in large area units

In large khesegs or bags that were selected more that once by the standard PPS procedure, the corresponding number of clusters were selected in the final stage. (For instance, if a unit was selected twice in the first stage, two clusters were selected in the

[3] This resulted in the allocation of 1 to 4 soums per aimag.

unit in the second stage.) This happened seldom in Ulaanbaatar but often in the rest of the country. The selection procedure was as follows:

- In non-herder units, all households in all required clusters were selected in a single step – with unequal probabilities as described above – and then allocated to specific clusters by systematic, equal probability sampling.

- In herder bags, the starting points of the required clusters were equally spaced in the list of households.

Allocation of the sample in time

The 360 clusters selected in Ulaanbaatar, and the 12 or 24 clusters selected in each aimag center or rural soum were randomly allocated into the 12 months of survey fieldwork. The survey will thus visit a random sub-sample of 104 clusters (936 households) each month.

Description of the sample

Figure 1 below gives the total number of households in the population, the number of area units and the number of households of the proposed design, by stratum and by aimag.

Figure 1: Mongolia Socioeconomic Survey 2007
Number of households in the population,
and number of soums, clusters and households in the sample
by major stratum and by aimag.

Aimag		Total No. of households				Selected rural soums	Selected clusters				Selected households			
		Ulaan baatar	Aimag Centers	Rural Areas	Total		Ulaan baatar	Aimag Centers	Rural Areas	Total	Ulaan baatar	Aimag Centers	Rural Areas	Total
1	Arhangai	Àðõàí ãàé		4,399	19,877	**24,276**	**4**		12	48	**60**		120	384	**504**
2	Bayan-Ulgii	Áàÿí -ª ëãëé		6,289	15,039	**21,328**	**3**		12	36	**48**		120	288	**408**
3	Bayanhongor	Áàÿí õî í ãî ð		6,433	14,502	**20,935**	**3**		12	36	**48**		120	288	**408**
4	Bulgan	Áóëãàí		3,081	11,935	**15,016**	**2**		12	24	**36**		120	192	**312**
5	Govi-Altai	Ãî âù-Àëòàé		4,666	10,807	**15,473**	**2**		12	24	**36**		120	192	**312**
6	Dornogovi	Äî ðí î ãî âü		5,121	8,847	**13,968**	**2**		12	24	**36**		120	192	**312**
7	Dornod	Äî ðí î à		9,603	8,484	**18,087**	**2**		12	24	**36**		120	192	**312**
8	Dundgovi	Äóí äãî âü		3,429	9,199	**12,628**	**2**		12	24	**36**		120	192	**312**
9	Zavhan	Çàâõàí		4,074	15,855	**19,929**	**3**		12	36	**48**		120	288	**408**
10	Uvurhangai	ª âº ðõàí ãàé		5,566	23,227	**28,793**	**5**		12	60	**72**		120	480	**600**
11	Umnugovi	ª ì í º ãî âü		4,358	8,440	**12,798**	**2**		12	24	**36**		120	192	**312**
12	Suhbaatar	Ñ¿õáààòàð		3,510	9,829	**13,339**	**2**		12	24	**36**		120	192	**312**
13	Selenge	Ñýëýí ãý		4,691	17,502	**22,193**	**3**		12	36	**48**		120	288	**408**
14	Tuv	Ò³â		3,657	19,652	**23,309**	**4**		12	48	**60**		120	384	**504**
15	Uvs	Óâñ		6,345	13,455	**19,800**	**3**		12	36	**48**		120	288	**408**
16	Hovd	Õîâä		6,675	12,803	**19,478**	**2**		12	24	**36**		120	192	**312**
17	Huvsgul	Õ³âñã³ë		8,672	22,111	**30,783**	**4**		12	48	**60**		120	384	**504**
18	Hentii	Õýí òèé		4,545	13,396	**17,941**	**3**		12	36	**48**		120	288	**408**
19	Darhan-Uul	Áàðõàí -Óóë		18,666	4,075	**22,741**	**1**		24	12	**36**		240	96	**336**
20	Ulaanbaatar	Óëààí áààòàð	216,342			**216,342**		360			**360**	3,600			**3,600**
21	Orhon	Ï ðõî í		20,059	811	**20,870**	**0**		24	0	**24**		240	0	**240**
22	Govisumber	Ãî âüñ¿ì áýð		2,301	944	**3,245**	**0**		0	0	**0**		0	0	**0**
Total		**216,342**	**136,140**	**260,790**	**613,272**	**52**	**360**	**264**	**624**	**1,248**	**3,600**	**2,640**	**4,992**	**11,232**

Justification of the sampling design

The SES sampling strategy updates the design adopted by the Household Income and Expenditure Survey (HIES) in 2001, on the basis of the 2000 census.[4] The amendments take three major factors into consideration:

- The availability of a more recent sample frame, developed by the NSO on the basis of 2005 population figures from the local registration agencies. These figures show major changes in the population distribution over the past five years (Ulaanbaatar, for instance, has grown from to 755,000 to 912,000 in the period,) revealing that progressively larger parts of the population were escaping from the scope of the surveys based on the 2001 census.

- A raising concern for measuring and understanding the economic importance of the self-employed (often but wrongly referred to as "the informal sector.") The SES design responds to this challenge by selecting households known to contain self-employed preferentially in the final sampling stage, and by marginally increasing the portion of the sample allocated to Ulaanbaatar, where most of the self-employment activities are known to be concentrated.

- The fact that the Socioeconomic Survey will need to be fielded with an operational strategy that is significantly different from the one used by the HIES so far. The proposed design recognizes that the integrated survey will need well-trained interviewers, organized into teams and devoted to the survey on a full-time basis, and that it will also need a decentralized data management component that integrates computer-based quality controls to fieldwork. Several teams will operate in Ulaanbaatar and one or two teams in each of the other aimags.

The SES sample conserves many features of the HIES, particularly a total sample size of 11,232 households per year. However, while the HIES recognized only four explicit strata ([1] Ulaanbaatar, [2] Aimag Capitals and Small Towns, [3a] Soum Centers and [3b] Countryside,) the SES collapses the small towns with the rural areas, and explicitly recognizes many more aimag-level strata. These enhancements are justified by various reasons:

- Neither the HIES nor the SES will be able to produce reliable aimag-level estimations. The only way of achieving this would be to significantly increase the total sample size, to around a thousand households per aimag – something that cannot be recommended at this moment. However, the allocation of at least a minimum sample to each aimag and the decentralization of fieldwork and data management are significant first steps in the direction of this ambitious goal.

- Small towns are indeed very small (most of them have less than 2,000 households) and collectively represent only about 10 percent of the population. Under these conditions their presence as partners of aimag

[4] For a description of the HIES design, see Levinson, A. *Mongolia Income and Expenditure Survey*. Ulaanbaatar, October 2000; and Muñoz, J. *Mongolia Household Survey System: Sampling Implementation and Survey Integration*. Ulaanbaatar, February 2001.

capitals in an explicit stratum was not compatible with a minimum urban size in each aimag, and with the deployment of at least one interviewer in each aimag capital. It is important to underline that the SES design does not exclude small towns from the sample – they are just transferred to another stratum. This may require some care at the analytic stage, but does not engage the comparability of the SES with the HIES series.

Another deviation of the SES relative to HIES is that, although the total size of the rural sample (4,992 households) remains unchanged, it will now consist of 96 households (12 bags) in each of 52 soums, rather than of 64 households (8 bags) in each of 78 soums. This probably will bring about slightly higher cluster effects, but it will also improve the capture of seasonal variations, and will keep rural interviewers busy year round at a marginal cost.

A final difference between the SES and HIES samples is in the selection technique adopted for the herder communities: whereas the HIES selected both herder and sedentary clusters by systematic sampling within their respective bags, the SES will try to facilitate fieldwork selecting herders by circular cluster sampling instead – the underlying assumption being that households that appear close to each other in the lists are also likely to be neighbors in the field. Cluster sampling may bring about larger sampling errors than systematic sampling, but it is also expected to be less vulnerable to nonresponse and to the selection biases that have affected the recent rounds of the HIES.

Implementation of the sampling stages

Initial sampling stages

In December 2006, the 2005 population registration figures were organized into adequate sample frames for the three major strata, and the first sampling stage was conducted in Ulaanbaatar and the rural areas. This resulted in the selection of 357 khesegs in the capital (three of them selected twice,) and 52 rural soums (allocated into aimags as shown in Figure 1.) A household listing operation was conducted in all of these khesegs and soums.

The first sampling stage was also conducted in the aimag centers, but the relatively small total number of bags per center resulted in almost all of them being chosen.[5] Under these conditions, it would have been easier and better to ignore this preliminary exercise and simply conduct the household listing operation in all urban centers, leaving the actual allocation of clusters into bags for a later moment, when updated figures would be available for all of them. This was actually done in all but four of the aimags: Bayan-Ulgii, Govi-Altai, Tuv and Uvs. In each of these aimags, the listing operation took place in all but one of the bags, thus making the December exercise a *de facto* zero-th sampling stage that will require a very small adjustment in one of the estimation formulas presented below (Formula 2.)

[5] There are 168 bags in total in all aimag centers (Ulaanbaatar and Govisumber excluded.) All but seven of them were selected in December, and one half of those selected were selected more than once.

Household listing operation

The household listing operation was conducted in May and June of 2007 in all selected khesegs of Ulaanbaatar, all 52 selected rural soums, and almost all aimag centers, as explained above. The final database contains over 275,000 households in total and, beyond its immediate utilization for the SES, has a high potential value as a master sample frame for other household surveys conducted by the NSO in future years.

A sudden rescheduling of the survey launching date prevented the implementation of uniform practices for the computerization of the household lists, and resulted in a wide variety of heterogeneous spreadsheets demanding considerable data management effort in order to build from them a reliable frame for the subsequent sampling stages. One of the aimags (Hentii) delivered its files only days before the launching of the survey, and in some of its bags the lists consisted of simple sequences of serial numbers, without household addresses or names.

The operation was also confused in one of the Ulaanbaatar districts (Baianzurj,) where the maps of four horoos had been redrawn and local staff had trouble identifying the boundaries of the twenty "old" khesegs that had been selected in these horoos. This was solved by taking a sample of twenty "new" khesegs in the affected horoos, using the most recent population figures for the PPS selection. The December 2006 exercise can also be considered in this case as a *de facto* zero-th sampling stage demanding a small adjustment in Formula 1 below.

Some aimags or districts delivered lists of a few extra khesegs or bags, in addition to those they had been asked for. These additional units were ignored in the subsequent sampling stages.

Selection of clusters

The final allocation of clusters into bags and the selection of the target and reserve households in each cluster were conducted with a special-purpose program (an Excel macro.) Additional programs were responsible for the production of the two associated survey instruments:

- The ***Cluster Tracking Sheets*** will be used by fieldworkers to identify the selected households. They give the name of the household head, the household size and the address of the 8 (or 10) target households and 3 reserve households in a cluster.

- The ***Household Listing Printouts*** reproduce the list of all households in the area, flagging the selected households among them. These lists will facilitate the identification of the selected households when precise addresses are not available.

Figures 2 and 3 below reproduce one Cluster Tracking Sheet and a page of the corresponding Household Listing Printout. Notice that although in sedentary areas households known to contain self-employed are preferentially chosen, neither of these instruments reveal this information to the fieldworkers. They have been instructed to probe for <u>all</u> self-employment activities of <u>all</u> members in <u>all</u> households, not just in some of them.

Figure 2: A Cluster Tracking Sheet
(50 percent of actual size)

Í èéãýì ýäèéí çàñãèéí ñóäàëãàà 2007
Ò¿¿âýðëýãäñýí í ýãæèéí ì àÿãò

Ò¿¿âðèéí í ýãæèéí äóãààð	Ñàð
1143	8

ÕÎ Ò

Àéì àã	20	Óëààí áààòàð
Ä¿¿ðýã	7	Óá - ×èí ãýëòýé
Õî ðî î	18	Óá - ×èí ãýëòýé - 18 õî ðî î
Õýñýã	09	Óá - ×èí ãýëòýé - 18 õî ðî î - 9

ª ðõèéí äóãààð	Æààñààëòûí äóãààð	ª ðõèéí òýã¿¿ëýã÷èéí í ýð	Àì á¿ë	Õàÿã
Ñî í ãî ãäñî í º ðõ				
01	3	М.Хүрэлбаатар	4	Я-37-151
02	16	С.Дулам	3	Я-37-142
03	41	Д.Энх-Амгалан	4	Я-37-573
04	55	Д.Ганбаатар	4	Я-38-580
05	66	М.Дашням	5	Я-38-588
06	92	Л.Гантайван	3	Я-42-692
07	103	Г.Дорж[illegible]	3	Я-42-702
08	118	Ш.Мөнхбаатар	7	Я-40-603б
09	147	Г.Цэрэннадмид	7	Я-40-624
10	160	Ю.Тосмаа	5	Я-43-709

Í ºº ö º ðõ

¯ÑÃ-ûí ì ýðãýæèëòýí í ºº ö º ðõèéã àø èãëàõ ýðõèéãî ëãî í î .
Í ºº ö º ðõèéã àø èãëàõ áî ëî í ø àëòãààí û ãí àðèéí òî ãî ðî î éëí î .

11	29	Г.Цэрэндолгор		4	Я-37-563
		¯í äñýí º ðõèéí äóãààð:		Øàëòãààí	
12	79	М.Далантай		1	Я-38-598
		¯í äñýí º ðõèéí äóãààð:		Øàëòãààí	
13	133	Ю.Үйэрг		7	Я-40-615
		¯í äñýí º ðõèéí äóãààð:		Øàëòãààí	

Figure 3: Page of a Household Listing Printout
(50 percent of actual size)

ÆÄ	ÒÏ Ä	ª Ä	Àì á¿ë	Í ýð	Õàÿã	Òàéëáàð
1			4	П.Борис	Я-37-152	
2			4	М.Цолмон	Я-37-151	
3	1143	01	4	М.Хүрэлбаатар	Я-37-151	
4			3	Д.Нацаг	Я-37-150	
5			5	Н.Цэдэнбал	Я-37-149	
6			4	Л.Наранзул	Я-37-147	
7			5	Д.Равжих	Я-37-147	
8			3	О.Авирмэд	Я-37-146	
9			3	С.Нарангэрэл	Я-37-146	
10			3	В.Мөнхбаатар	Я-37-146	
11			4	Б.Гүрэвбат	Я-37-145	
12			6	П.Сергей	Я-37-143	
13			3	П.Валентина	Я-37-143	
14			8	Б.Батбаяр	Я-37-142	
15			3	О.Төмөрчөдөр	Я-37-142	
16	1143	02	3	С.Дулам	Я-37-142	
17			7	Н.Даваас	Я-37-141	
18			2	Т.Мэнсэл	Я-37-140	
19			5	Т.Ганболд	Я-37-139	
20			3	Ч.Гармаа	Я-37-139	
21			4	Б.Ганбаатар	Я-37-139	
22			4	Б.Ганболд	Я-37-139	
23			5	Д.Отгон	Я-37-139	
24			5	Ч.Ганболд	Я-37-558	
25			4	Х.Энхбаатар	Я-37-558	
26			3	Ц.Донид	Я-37-559	
27			2	Д.Баяраа	Я-37-560	
28			4	Ч.Коля	Я-37-563	
29	1143	11	4	Г.Цэрэндолгор	Я-37-563	
30			5	Г.Аюурзана	Я-37-563	
31			3	Ш.Цэрэннадмид	Я-37-563	
32			5	Ц.Биндэръяа	Я-37-564	
33			6	Н.Жамъянбаатар	Я-37-565	
34			3	Х.Энхбаяр	Я-37-566	
35			6	Ч.Мөнхбаяр	Я-37-567	
36			7	О.Халиухай	Я-37-568	
37			6	Г.Эрдэнэцогт	Я-37-569	
38			3	Я.Дэлгэрмаа	Я-37-571	
39			4	Д.Тогтох	Я-37-572	
40			1	Л.Шүрх	Я-37-574	
41	1143	03	4	Д.Энх-Амгалан	Я-37-573	
42			7	С.Эрдэнэ	Я-37-575	
43			2	Ц.Энхтуяа	Я-37-575	
44			3	Б.Батсайхан	Я-37-576	
45			4	Ж.Чулуунням	Я-37-01	
46			4	П.Бат-Эрдэнэ	Я-37-02	
47			2	Ч.Өмдэрэлмаа	Я-37-03	
48			3	Д.Наранцэцэг	Я-37-06	
49			8	У.Оюунчимэг	Я-37-07	
50			2	У.Оюунцэцэг	Я-37-07	
51			1	Д.Лонжид	Я-38-578	
52			2	Ш.Нанзад	Я-38-579	
53			4	Н.Дашням	Я-38-579	
54			3	Д.Өнөржаргал	Я-38-579	
55	1143	04	4	Д.Ганбаатар	Я-38-580	

Selection probabilities and sampling weights

Ulaanbaatar

In Ulaanbaatar, the selection probability of household ij in kheseg i is given by

$$p_{ij} = \frac{360\ n_i}{216{,}342} \times \frac{10\mu_{ij}}{\sum_{\alpha=1}^{n'_i} \mu_{i\alpha}} = \frac{3{,}600\ n_i \mu_{ij}}{216{,}342 \sum_{\alpha=1}^{n'_i} \mu_{i\alpha}} \quad \text{.....(1)}$$

where

360 is the number of clusters selected in Ulaanbaatar;

n_i is the number of households in kheseg i, as recorded in the 2005 administrative registration files used as a sample frame for the first sampling stage;

216,342 is the number of households in Ulaanbaatar, as recorded in the first stage sample frame (see Figure 1;)

10 is the number of households per cluster;

n'_i is the number of households in kheseg i, as reported by the 2007 household listing operation;

$\mu_{i\alpha}$ is the measure of size assigned to household $i\alpha$ in the second sampling stage: $\mu_{i\alpha} = 2$ for households assumed to contain self-employed and $\mu_{i\alpha} = 1$ for the rest of the households.

Aimag Centers

In the aimag centers, the selection probability of household hij in bag (or sub-bag) hi of aimag h is given by

$$p_{hij} = \frac{b_h n_{hi}}{N_h} \times \frac{10\mu_{hij}}{\sum_{\alpha=1}^{n_{hi}} \mu_{hi\alpha}} \quad \text{.....(2)}$$

where

b_h is the number of clusters selected in the aimag center: $b_h = 24$ in Darhan-Uul and Orhon and $b_h = 12$ in the other aimag centers;

n_{hi} is the number of households in bag hi, as recorded in the 2007 household listing operation;

N_h is the number of households in the aimag center, as reported by the household listing operation;

10 is the number of households per cluster;

$\mu_{i\alpha}$ is the measure of size assigned to household $i\alpha$ in the second sampling stage: $\mu_{i\alpha} = 2$ for households assumed to contain self-employed and $\mu_{i\alpha} = 1$ for the rest of the households.

Rural Areas

In rural areas, the selection probability of household hij in bag (or sub-bag) hi of soum h is given by

$$p_{hij} = \frac{52\ N_h}{260{,}790} \times \frac{12\ n'_{hi}}{N'_h} \times \frac{8}{n'_{hi}} = \frac{4{,}992}{260{,}790} \times \frac{N_h}{N'_h} \quad \text{.....(3)}$$

where

52 is the total number of rural soums selected;

N_h is the number of households in soum h, as recorded in the 2005 administrative registration files used as a sample frame for the first sampling stage;

260,790 is the total number of rural households, as recorded in the first stage sample frame (see Figure 1;)

12 is the number of clusters selected in soum h;

n'_{hi} is the number of households in bag hi, as recorded in the 2007 household listing operation;

N'_h is the number of households in soum h, as reported by the household listing operation; and

8 is the number of households per cluster.