SDA July 89 | Oct 89

Copy

THE GAMBIA SDA PROJECT

SAMPLE DESIGN

TECHNICAL REPORT

L.M. ASSELIN
SENIOR STATISTICIAN
AF1SD
JANUARY 1990

THE GAMBIA PROJECT

SAMPLE DESIGN-TECHNICAL REPORT

1. Preliminaries.

1.1 The objective of this report is to write down any technical information associated with the process of designing the sample for a SDA survey in The Gambia, process which has taken place between July 89 and Oct 89.

1.2 First, it is important to note that in this interval, the statistical program for the first year has been modified. In July 89, it was envisaged to implement a Household Integrated Survey (HIS), with the questionnaire which was then presented. But the methodology of the SDA Priority Household Survey having received an important development in Aug-Sep 89, it then appeared that this type of survey would be more appropriate for the Gambia project in its first year. It was then proposed, and accepted in Oct 89, that a PHS be conducted in the first year.

1.3 For the sample design, a first mission was made by L.M. ASSELIN in July 89, to assess the local ressources available for statistical work, to identify and, if necessary, to begin to build a sampling frame for a household survey which was then planned to be a HIS.

1.4 A second mission was made in Oct. 89 by L.M. ASSELIN to present the proposal of the PHS and to help in the finalization of the sampling frame, in the sample design and in the selection of the sample at the first level.

2. Sampling frame: existing lists.

2.1 Census 83 as a starting point.

A) The Gambia 1983 Population Census appeared immediately as the primary information basis for providing a list of sufficiently small units with a population size. Results of this census have recently been published (1987-89) by the Directorate of Statistics. Annex 1 gives a summary of the results, with the information relevant for planning a survey: total population, number of compounds and number of households by district. We must recall that The Gambia is divided in 6 divisions, 8 local governments areas (LGA) and 35 districts. For the census, the LGA of Banjul was divided in 4 "districts", and the LGA Kanifing in 9 "districts". For these last 9 "districts", the number of households and of compounds has not been published, so that it was estimated by using the mean population per compound and per household for Kanifing. A map is given in annex 2 for identification of divisions and LGA.

B) Census data are on a Wang mainframe. In July, the Directorate of Statistics had in hands an unique list of Enumeration Areas (EAs), a computer print out giving the total population of each EA by LGA and district, with the EA 5-digit code: DD-SSS, where DD is the district code, and SSS is a sequential number. To help in using this listing, there is a document called "Enumeration Area List", giving for each settlement (village, urban neigbourhood) the number of EAs and their codes. When an EA covers more than one settlement, the list of settlement names is given. There does not exist any list giving, at the EA level, the number of households or of compounds. There are 1304 EAs, with a mean population size of 527.

C) From the census data, another print out has been produced, giving, for each settlement, the distribution of population by sex and age-groups. This list is being entered in the unique micro-computer belonging to the CSD, a PC-XT. The list is entered as a Lotus file.

D) Census maps, giving boundaries of EAs, are available even if some of them are quite damaged. The scale is usually 1/5000 in urban areas (with identification of compounds), and 1/50000 in rural areas. Due in part to the filing equipment available, they are not easy to retrieve, and some of them seem to be missing, since they could not be found in the course of building a sampling frame.

For the whole country, the best map we found to identify settlements is the road map published by the Government of the United Kingdom in 1980, with a scale of 1/250000. It is available at the Survey Department in Banjul.

2.2 Other possible lists.

From our consultation with the CSD, we conclude that no other complete list of population units exists at a sufficiently disaggregated level, with a size estimate. Contrary to other countries, we do not find in The Gambia a kind of administrative record with a regularly updated population information at the settlement level. Nevertheless, through the discussion, we learnt that for taxes purposes, information should exist, at the local administration level, on the number and type of building structures, for each compound. Also, at the end of the second mission, we were informed that the Department of Planning of the Ministry of Agriculture, for its own annual survey needs, has built in 1988-89 a list of rural settlements, with the number of "dabadas". This list exists as a micro-computer file. This deserves certainly more attention, but the facts that the basic unit is the settlement (variance due to size, problem of identification in the field), that the measure of size is the number of dabadas(concept very specific to agricultural sector, difficult to update), and that the urban area is not covered, are important limitations to

its use as a frame for more general uses like SDA surveys.

3. The rural-urban distinction.

For the SDA analysis, the rural-urban distinction is a very important one. For field work organization, it is also necessary to distinguish between areas of high population density ("urban" area) and those of scattered population ("rural" area). Also, inter censuses population changes (growth or decrease) are different in rural and urban areas, and accordingly, the methodology of updating the EAs list. In The Gambia, there is no official definition of urban area. Even in the 1983 census, such a classification was not used for publication of results. Implicitly, this means that in in this small country, the urban area is an ambiguous reality.

In July, a long discussion was held in the CSD in view of an operational definition of rural and urban areas. Among the criteria for being an urban area were mentioned: an important number of wage earners, important commercial activity, physical infrastructure. Finally, the working group concluded on the following approach:

a)first, retain a size criterion: list the settlements having a 1983 population of 2500 or more. This gave 21 settlements.

b)from the list so obtained, exclude settlements where agriculture is the dominant activity and where commercial activity is not important: 11 settlements were then eliminated.

c)add settlements whose population is close to the size limit, but which are predominantly non agricultural. Only one settlement was so added on this basis, Kuntaur Wharf Town (pop 2059).

d)the final list for urban area is:

TABLE 1

DEFINITION OF URBAN AREA IN THE GAMBIA

Division	Settlement	1983 pop.
Banjul-Kombo St. Mary	Banjul	44188
	Kombo St.Mary	101504
Western Division	Brikama	19644
	Lamin	5799
Lower River Division	Soma	4789
North Bank Division	Farafeni	10168
Maccarthy Island	Kaur	5149
	Kuntaur Wharf Town	2059
	Georgetown	3068
	Bansang	3964
Upper River Division	Basse	9241

This gives a total 1983 urban population of 209573, i.e 30% of population (687817).

Everybody was aware of some unavoidable arbitrariness in this definition of urban area. But a decision was to be made, and there was a consensus among the working group on the list finally retained. It should be mentioned that except for Banjul and Kombo St.Mary, the most appropriate term would probably be "semi-urban" instead of "urban".

As defined, the rural-urban factor will be the first explicit stratification criterion.

4. Other stratification criteria.

In the SDA methodology, stratification should focus on on the issue of identifying socio-economic (SE) groups, these groups including vulnerable groups, poverty groups and target groups. Most of the criteria which can be suggested for the identification of these groups are at the household level (see annex 3). At the primary sampling stage (PSU level), within the rural area, we could think first of agro-ecological zones. Due to the smallness of the country, to its shape (a long stripe from west to east along the Gambia river), there are not strongly identified agro-ecological zones. A criterion which was raised to our attention is the distinction between lowland and highland villages, according to which soil cultivated are different and more or less favourable to rice growing. But this criterion cannot be easily introduced for all EAs, and we suggest that it be observed on selected PSUs. Second, there is the distance to markets criterion. Again, this is not know at the EA level, but a study should be made to see if the 35 administrative districts cannot be classified according to their degree of accessibility. If so, a geographical stratification according to this classification of districts would be relevant for SDA analysis.

In the urban area, it could be interesting to make a distinction between the capital area (Banjul Kombo St. Mary), and other urban area which is more or less semi-urban. But this should not be considered here as a very important criterion. Type of urban neigbourhood, according to mean quality of dwelling, public infrastructure, etc., is another criterion which can be mentioned, but it would be more relevant in a more urban country. Also, this criterion is not available at the EA level and would require a supplementary field operation to be retained as a stratification variable. If considered relevant, it could be measured in a community questionnaire for selected urban EAs.

5. Sampling frame: the need to update the 1983 EA list.

5.1 Obviously, the population size of EAs has changed

considerably in some areas since 1983, and some kind of updating operation is needed for variance control and field work organization. The various types of change were discussed at length. It came out that the most important and concrete effect of these various changes was the physical extension (new compounds) at the outskirt of some settlements. This growth by extension can modify seriously the relative size of EAs, and it was recognized as the point on which should focus the updating of the 1983 list. On this basis, a first distinction was introduced: settlements having 2 or more EAs, which were called settlements of type A, and settlements being covered by only one EA, or part of an EA, called settlements of type B. Similarly, EAs belonging to A-settlements are called EAs of type A, and an EA covering one or more settlements is called an EA of type B. The risk of significant physical extension was considered to be essentially confined in A-settlements, and even not in all of them. Urban settlements are obviously included in A-settlements and are identified as A1, the other ones as A2-settlements.

5.2 Different kinds of updating operations were considered, all being a mix of direct observation and regional meetings with local representatives. It was first planned that these operations be made before any sample selection, at least in Banjul Kombo St.Mary. Due to the lack of physical resources (vehicles), it was decided to update only a sample of large primary sampling units, these being A1 settlements, large A2 settlements, or parts of these settlements. But even for Banjul itself, the CSD came finally to the conclusion that growth by extension was not significant, this growth in the capital having occurred essentially in Kombo St.Mary. Banjul was then excluded from the updating operation. So in some areas, this sampling operation will be the first of a three-stages sample design. The definition of these primary sampling units (PSUs), called sectors, was made from the census maps, on the basic principle to try to define PSUs having approximately the same size, for variance minization in a subsequent equal probability 2nd stage sampling. The approach was to subdivide large A1 and A2 settlements on the basis of 1983 EAs boundaries, using the 1983 population and any other source of information on areas where extension took place.

5.3 In urban area, where the probability of extension is higher, the target size (total pop.) for the sectors was fixed at 5000. In rural area, only A2 settlements with a 1983 population of 2000 or more were considered as having a significant probability of extension; the set of these settlements was called the rural dense area. It should be noted that the 11 settlements already discarded (see para 3b) in defining the urban area belong to this rural dense area. As many of those settlements had a size between 2000 and 3000, it was decided that large A2 settlements would be divided in sectors with a target size of about 2500. The list of sectors , including those built by division of large A1 and A2 settlements, is given in annex 4 with the EAs codes. The total

number of urban sectors is 36, and of rural sectors, 29.

5.4 The updating operation itself, in sectors which were to be eventually selected, was defined as follows:

a) in the field, with old census maps in hands, define new EAs having approximately the same size, about 500 persons.
b) this will be made by counting the number of inhabited compounds, with a target number of compounds determined in each sector on the basis of 1983 average compound size in the relevant district (see annex 1): e.g., in a district where average compound size is 13.2 persons, the number of inhabited compounds per new EA should be approximately 38.

Those new EAs will then become the secondary sampling units (SSU) in these updated areas. Maps will also be updated by reporting the main landmarks required to identify the new EAs boundaries.

6. The sampling frame building.

6.1 The list of EAs described in para 2.1 is not ordered as required by the sample design, and is not computerized. It appeared in July that that a new list, more suitable to the needs, could not be produced quickly from the mainframe. Taking into account the small number of EAs (1304) and the need to get immediately a list giving the EAs according to type (A1,A2,B), a new list was manually built from the print out. It was a one-day operation, for a team of 6 persons.

6.2 The need to transform and to complete this list in various ways brought the decision to computerize it on micro-computer, as a LOTUS file. This work was done by the SDA unit in Washington. The list was then transformed during the second mission by different operations:

a) building of sectors in urban and dense rural areas (aggregation of EAs).
b) estimation of the number of households per EA, on the basis of mean household size per district.
c) merging of EAs with a 1983 population of less than 250 with a larger EA. This operation was made in view of insuring that there is a sufficient number of households in each EA for different types of SDA surveys, and also with an objective of reducing variance due to size, especially in case equal probability sampling would be used. This work required to go back to census maps, to identify the most appropriate adjacent EA to which merge the small EA. About 60 EAs were then merged, in areas not retained for the updating operation.
d) splitting of EAs with a population larger than 1000, to reduce the field work and also the variance due to size (for eq. pr. sampling). About 15 EAs were so splitted in "north" and "south"

parts, the precise delimitation to be completed in the field if the EA part is selected.

6.3 The sampling frame is given in annex 5, beginning with the sectors followed by the EAs, by rural-urban,LGA and district, with:

a) a sequential number for urban and rural sectors, and also for urban and rural EAs.

b) the type of unit:
- A11: in Banjul Kombo St.Mary ("strictly" urban)
- A12: in other urban.
- A21: in A2-settlements (more than 1 EA) with pop>=2000.
- A22: in A2-settlements with pop<2000.
- B : from B-settlements (1 or more settlements in the EA).

c) a minimum identification of the settlement to wich it belongs, except for EAs of type B. The complete identification can be found in the "ENUMERATION AREA LIST".

d) LGA code

e) District code.

f) EA code for EAs, and # of EAs 1983 for sectors.

g) pop 83.

h) estimated # of hlds. in 1983.

i) for EAs, the cumulative pop. size, for rural and urban, in view of easing eventual sampling with pps.

It should be noted that this frame was completed by a stratum code and ordered differently for the specific needs of the PHS (see section 8). In annex 5, it is given in a format which could be useful for general purposes.

7. Sample design of a Priority Household Survey in the first year.

7.1 Basic parameters.

a) The objective of a PHS in the first year is to get as soon as possible some basic information to monitor the social effects of the adjustment program, and particularly, to identify the most vulnerable groups.

b) The same resources planned for the IHS are to be used for the PHS: 5 mobile teams, 1 for Banjul Kombo St.Mary, the other four being based in four regional antennas to be created: Brikama, Mansa Konko, Georgetown and Basse. Each team has 2 enumerators, a controller, a driver and a data-entry clerk.

c) The survey is to be completed in a short period of time. Resources available are then determinant for the total sample size. The daily work load for one enumerator is estimated to 4 interviews. In a 5 working days week, 200 interviews can be completed by the 5 teams. The total sample size was then fixed to 2000 households, which, in addition to the listing time, would require 10 weeks of field work. The length of the listing period depends on the number of selected EAs, i.e. on the number of selected households per EA. The objective was to have a large number of selected EAs, in view of identifying a large number of

SE-groups while reducing the variance due to the 1st stage of selection (1st and 2nd stages in updated areas). To allow quick preliminary results from the first 1000 households, it was decided to have two consecutive rounds of 1000 households, each round including two steps: households listing in selected EAs, and interviews.

7.2 Sample structure.

a) stratification and sampling methods.

As explained in paras 3 and 4, the only explicit stratification criterion retained for analysis needs is the rural-urban factor. But stratification was also introduced for operational and precision purposes.

In the urban area, Banjul Kombo St.Mary, as the capital region and as the specific territory of one of the 5 teams, was identified as a specific stratum, A11. Other urban area is then the complementary stratum, A12. Banjul Kombo St.Mary was again subdivided in two sub-strata, with a two-stages sample in Banjul (A11B), and a three-stages sample in Kombo St.Mary (A11K).

In the rural area, we have to distinguish the dense rural area (A21), where a three-stages design is to be used, and the strictly rural area (B), where EAs are the PSUs. The strictly rural area is made of all B-settlements and all A2-settlements whose size is less than 2000 persons.

In all strata with a three-stages sample (A11K,A12 and A21), equal probability sampling will be used at the first stage, since the sectors have been built to be now approximately of the same size. In the selected sectors, new EAs will again be defined so that they have approximately the same size (500), and then, at the 2nd stage, equal probability sampling will be used.

In strata with a two-stages sample (A11B and B), the situation is different. For controlling variance due to EA size, a pps procedure would be used if we were in a context of an unique, isolated, one-shot survey. But here the survey program will extend over four years, and we are in a dynamic situation in the sense that the program is not fully determined. A PHS is a natural candidate for repetition. When will it be repeated? Will it be with a partial replacement of selected PSUs? With the same sample size? For other SDA surveys (community survey, integrated survey), eventually introduced in the program, it will be interesting to look at integration and consistency with preceding surveys. So, to insure a very flexible sample design in the context of a not well determined system of surveys, it was decided to control variance by stratification of EAs according to size, instead of using the pps procedure; equal probability sampling will then be used in size sub-strata.

In Banjul (A11B), with only 79 EAs, three sub-strata were defined:

A11BX: EAs with 83 population 250-424
A11BY: EAs " " " 425-649
A11BZ: EAs " " " 650-1000.

In the strictly rural area (B), with 798 EAs, five sub-strata were defined:

S1:	EAs	with	83 population	250-324
S2:	EAs	"	"	325-424
S3:	EAs	"	"	425-549
S4:	EAs	"	"	550-724
S5:	EAs	"	"	725-1000.

The cutting points in the size range 250-1000 were determined so that the ratio between two consecutive points be approximately constant, i.e that variation in the weight between the smallest and the largest unit be as small as possible in each sub-stratum. But compromises had to be made to take into account the distribution of EAs according to size, e.g. for the first cutting point in Banjul.

b) Allocation of the total sample to strata.

To allocate the total sample of 2000 households, the 1983 population of each stratum was computed (see annex 6). First, with a team to work exclusively in the capital region, the sample size for this strictly urban division was automatically fixed at 400 households. It was then allocated between the substrata (A11B-X-Y-Z and A11K) according to their size. The 1600 remaining households were allocated to the remaining strata according to their size. In each stratum, the sample was equally divided between the two consecutive rounds. In this allocation, numbers had to be rounded to take into account the number of households in each selected EA (see para c). Results of the allocation are given in annex 6. We can see that the sample size in urban area is 592 hlds., and is 1408 in rural area.

c) Allocation of sample between the 2 or 3 stages.

The optimal allocation of sample between the different stages is a complex operation, which requires cost and variance data and which must take into account practical considerations related to field work. The basic point is to determine the number of households to be selected in each selected EA. As mentioned in para 7.1 c), it was intended, in the PHS, to have a large number of EAs in the sample. The sample would then be well spread over the country, allowing a good geographical identification of vulnerable groups while reducing the variance at the EA (and sector) level. The perspective for a PHS is to be repeated in a sufficiently short period so that the same PSUs can again be used with the same listing, or with a light updating of the household list. Then, even if in each survey the number of households per PSU is small, the whole take would be increase in a sequence of surveys.

For practical field work organization, the minimal sample size in a rural EA would be the daily work load of one team, i.e. 8 households, and in urban area, the daily work load of one enumerator, i.e. 4 households. This was the starting basis.

In Banjul (A11B), the number of 4 hlds./EA was retained. In the strictly rural area (S), it was decided to retain 8

hlds./EA. In Kombo St.Mary (A11K) and other urban area (A12), with a three-stages design, it was not possible to keep 4 hlds./EA: the number of EAs and of sectors to be selected would have been so large that it would have been no more significant to update only the selected sectors, or to list households only in the selected EAs. Then the number of hlds./EA was also taken as 8 in those urban strata. Finally, in the dense rural stratum (A21), again with a three-stages design, 8 hlds. were to be selected in each EA.

In strata with a three-stages design (A11K,A12,A21), all possible combinations of number of sectors and number of EAs were considered to give the targeted number of households. In fact, taking into account the constraint of using multiples of 8, the combination finally retained in each stratum came out almost immediately, to keep the advantages of the 3-stages design, i.e to reduce the field work in the updating and listing operations. In urban sectors (A11K and A12), 3 EAs will be selected in each selected and updated sector. In rural sectors (A21), smaller than urban sectors, 2 EAs will be selected in each selected sector.

The detailed results of the sample allocation are given in annex 6. To sum up, 17 sectors will have to be updated for each round, and 132 EAs will have to be listed in each round (44 in urban, 88 in rural). The team in Banjul Kombo St.Mary will have to list 32 EAs, and each regional team, 25 EAs, for each round. The Banjul team could be helped by the reserve team so that the listing time be approximately the same for every team.

8. Sample selection at the first level.

A last operation was made on the sampling frame before proceeding to the selection of PSUs. The stratum code was introduced in the file for each PSU. The list was sorted by strata, and within each stratum, by LGA and district.

The sample was selected at the 1st level (sectors, EAs) for the whole survey, and randomly allocated to each round so that each sub-sample is representative of the whole country with the same probabilistic structure. In strata Kombo St.Mary and other urban (A12), with a high sampling rate, sectors were drawn by simple random sampling (eq. prob.), the first half of the sample being retained for the first round. In all other strata, Banjul, dense rural and strictly rural, equal probability systematic sampling was used, the 1st,3rd,etc. unit being for the first round. It is important to note that with the geographical ordering of the list, we have with systematic sampling an implicit stratification according to LGA and district (see para 4 for the interest of this implicit geographical stratification). While progressing in the particular case of each stratum, to keep equiprobability sampling and the advantage of implicit stratification, cyclical systematic sampling appeared as the best procedure, the most appropriate period being usually the smallest integer close to N/n. The list was completely read, and in case of a sample n' larger than n, (n'-n) selected units were eliminated at random. The selected PSUs are given in annex 7, for each stratum.

In each selected sector, EAs will be drawn with equal probability, and in each EA, households will also be selected with equal probability (systematic sampling).

9.The weighting procedure.

The sample is not self-weighted, and in fact, this property was never sought as an objective.

9.1 The weight in strata with a 3-stages design (A11K,A12,A21).
Notation:

S_h : total # of sectors in stratum h.

s_h : # of selected sectors in stratum h.

M_{hi} : total # of new EAs in selected sector (h,i).

m_h : # of selected EAs in any selected sector in stratum h.

N_{hij} : total # of hlds in selected EA j of sector (h,i).

n : # of selected hlds in each selected EA.

Then, the weight for every selected household in EA (h,i,j) is:

$$W_{hij} = \frac{S_h}{s_h} \cdot \frac{M_{hi}}{m_h} \cdot \frac{N_{hij}}{n}$$

Here, m_h = 3 in A11K and A12
= 2 in A21
n = 8 in all those strata.

9.2 The weight in strata with a 2-stages design (A11B-X-Y-Z, S-1-2-3-4-5)

Notation:

M_h : total # of EAs in stratum h.

m_h : # of selected EAs.

N_{hi} : total # of hlds. in selected EA i, in stratum h.

n_h : # of hlds selected in each selected EA in stratum h.

The weight for every selected hld. in EA (h,i) is:

$$W_{hi} = \frac{M_h}{m_h} \cdot \frac{N_{hi}}{n_h}$$

Here, n_h = 4 for A11B.
= 8 for S.

Table 2 gives the weight associated to the first sampling stage in every stratum.

TABLE 2

WEIGHT FOR Ist SAMPLING STAGE

Stratum code	# PSUs	# selected PSUs	Weight
A11BX	18	4	4.5
A11BY	35	12	2.92
A11BZ	26	12	2.17
A11K	22	12	1.83
A12	14	8	1.75
A21	29	14	2.0 (*)
S1	114	12	9.5
S2	150	20	7.5
S3	247	44	5.61
S4	208	48	4.33
S5	79	24	3.29

(*) standard systematic, with random start 2.