Case Study No. 8

DEVELOPING A SAMPLING STRATEGY FOR A NATION-WIDE SURVEY IN MALAWI

1. Introduction

This case study concerns one of several land utilisation studies that were funded by donor agencies in 1995 to help the Government of Malawi to develop a land policy and an actionable land reform programme, following Malawi's first multi-party elections in 1994.

Three studies were undertaken starting in 1995 to fill knowledge gaps:

- Public Land Utilisation Study (PLUS) funded by USAID;
- Customary Land Utilisation Study (CLUS) funded by EU; and
- Estate Land Utilisation Study (ELUS) funded by the UK Government Department for International Development (DFID).

Here we focus on the ELUS study to illustrate how the survey was set up with very little information concerning the population that was to be sampled. We discuss the way in which a sampling frame was set up for selecting estates and how this was used to develop a methodology for sampling estates.

2. Study objectives

In developing appropriate sampling procedures for ELUS, the study objectives and other key issues were:

- to obtain land use information at a national level, as well as separately for each of the three regions - North, Central and South;

- to obtain land use estimates separately for estates of different size categories;

- to obtain, from sampled estates, reliable estimates of land use and information on the socio-economic characteristics of the estates;

- to ensure procedures to be used were practically feasible;

- to ensure the sampling methodology was statistically sound. At the design stage, this required some approximate knowledge of variability in land areas in the different size categories so that sample size calculations were possible, while at the data analysis and reporting stage the methodology had to be such that levels of precision could be calculated for all key parameters.

3. Assessing relevant background information

Sampling aspects of the ELUS survey began with a review of existing literature and database information relating to the estate sub-sector. Discussions were held with the project land use specialist to determine sampling methodologies that were feasible in practice.

3.1 Tea and Sugar estates

Tea and sugar estates in Malawi have traditionally been large. They were well-organised and kept good records of their holdings. Sampling these estates therefore provided no special statistical challenge. In total there were 37 tea estates managed by nine tea companies. Five estates were selected at random to gather the relevant information.

There are only two sugar estates in Malawi. Both were visited and relevant information gathered via rapid appraisal techniques. Interviews were carried out with 10 workers from one estate and 12 from the other for additional information.

3.2 *Tobacco and other estates*

In the literature search, the most useful references found were reports of two relatively small surveys from 1990 and 1992. These and other smaller studies were limited to a few estates. Most used a case-study approach, and none was able to extrapolate results to a national level. There was also some doubt about the reliability of farm areas obtained since they were based on farmers' estimates.

Computerised sources of information were available at the Ministry of Agriculture (MoA) and at Auction Holdings sales records, but neither was found suitable. The former was highly unreliable, while the latter did not include estate areas, nor their exact locations.

It was therefore decided that the only viable option was to create a sampling frame specially for ELUS. This is not unusual in multi-stage surveys where the sampling frame is set up for the particular units selected at each hierarchical stage of sampling. However, in ELUS the situation was rather extreme because there was still no readily available information regarding the location of estates in any possible form of unit.

4. Developing a sampling strategy

The need to ensure accurate information on land use within selected estates suggested the use of air photos, and this in turn implied sampling estates according to their geographical distribution. Administrative units were of little assistance suggesting an area-based sampling procedure, in which one of the stages would involve aerial photography of selected grids of land, identified through an appropriate sampling procedure. In the end, DFID paid for aerial photography of the whole country – useful for other purposes as well as ELUS.

A more familiar use of satellite and aerial photographs is to look at land use by area with ground truthing to confirm image interpretation. Interestingly in this case, estate boundaries were identified on the ground and marked on the photograph before image interpretation within the boundaries, aided by some visit notes.

Several sampling protocols were discussed. Two schemes that appeared to satisfy the study objectives were presented for discussion at a planning workshop involving relevant stakeholders. The first involved a two-stage sampling design within the six Agricultural Development Divisions (ADDs) where most estates were concentrated. The second sampling scheme involved a three-stage design with districts as primary units. The latter proposal was favoured at the workshop since the first ignored estates that did not lie in areas of concentration of the estate sector. However, it was recognised that either methodology would run a high risk of not capturing enough large estates above 500 ha. A different sampling scheme was therefore adopted for estates above 500 ha, namely the use of the MoA database information to draw a random sample stratified by region. It was felt that the database information would be reasonably accurate for the very large estates since these were well-established over a long period of time.

This decision resulted in five size categories, i.e.

<20 ha 20 – <40 ha 40 - <100 ha 100 - <500 ha ≥500 ha.

Since each category was regarded as a separate population, the use of a different sampling scheme for large estates was not expected to have an adverse effect.

Sample size calculations for estate numbers were made using standard formulae for simple random samples. Precision levels were set to ensure that estate sizes, in the < 500 area size categories, were estimated to within 1 ha, 2 ha, 5 ha, and 30 ha of the true mean with 90% confidence. The standard deviations used were those derived approximately from information available in the Skills Gap Survey (ITAD, 1993).

5. The sampling protocol and its implementation

The sampling strategy for ELUS described above was arrived at after carefully examining other alternatives with a range of stakeholders, including the Land Utilisation Advisor and field personnel from the Ministry of Agriculture and the Ministry of Lands and Valuation. In the actual implementation of the protocol, decisions concerning the method of sampling had to be made in such a way that the procedures could be understood and appreciated by staff of relevant ministries and members of the Land Utilisation Studies Steering Committee.

The choice of districts as primary sampling units stemmed from two factors:

(a) Land use and land suitability estimates within each estate size category were needed at a regional and national level, with indications of the precision of such estimates; and

(b) The availability of approximate figures for estate numbers in each of the 24 districts. These were obtained with some difficulty from the MoA database since there was some resistance initially to making the information available to the ELUS team. The data covered only tobacco estates and were known to have 'ghost' and duplicate estates. However the information was thought adequate for the sample selection process.

Sampling of primary units was with probability proportional to the approximate number of estates so every district had a non-zero chance of selection. Thirteen districts were selected from a total of 24 districts to demonstrate adequate coverage (see Table 1).

Primary units were selected *with* replacement so that estimates of precision could be derived for key economic, social and land-use indicators. Derivation of precision estimates followed ideas presented by Rao (1975).

The second stage of sampling involved imposing a grid of 10 km × 10 km squares (blocks) on each selected district and choosing a pre-specified number of blocks from each district. The implementation process made use of 1:250,000 scale maps to exclude the possibility that areas considered for sampling would not include estates by virtue of being lake, swamp, national park, game reserve, forest reserve, etc. The remaining area gave a total of 586 blocks.

Subsets of the blocks were to form the second stage sampling units. The aim was to use these units to establish a sampling frame of estates within each block. The practicalities and time constraints involved in carrying out the estate listing process dictated that no more than 60 blocks could be included in the Listing Survey. We selected m_i blocks in district i in proportion to the number of blocks in that district. Table 1 shows the results, together with selection probabilities that resulted when districts were selected with replacement as primary sampling units.

Table 1. Number of blocks selected and selection probabilities for chosen districts.

Region	Sampled district	No. of blocks sampled m_i	No. of blocks in district M_i	Selection prob. of district p_i
North	Rumphi	2	15	0.2682
	Mzimba 1	9	90	0.6906
	Mzimba 2	9	90	0.6906
Central	Kasungu 1	5	51	0.3032
	Kasungu 2	5	51	0.3032
	Dowa	3	27	0.1203
	Lilongwe	6	53	0.2678
	Nkhotako:	2	16	0.0711
South	Mangochi1	4	43	0.5183
	Mangochi2	4	43	0.5183
	Machinga1	4	43	0.2525
	Machinga2	4	43	0.2525
	Zomba	2	21	0.1494
TOTAL		59	586	

The selection of blocks within a district by simple random sampling led to a more-or-less equal probability of selection for each of the 586 blocks. However, in the actual selection, a couple of blocks were disregarded because they were known to lie in particularly inaccessible territory. These were areas where travel by foot or by any vehicle could be very difficult along unmarked and poorly maintained dirt roads and would require a great deal of patience and perseverance to reach any destination. The lack of strict random sampling in this case was recognised as possibly introducing some bias. However the bias was felt to be of little importance relative to the near certainty of poor quality data due to field staff not providing an adequate coverage of the 100 km^2 area of such blocks.

The next step was to list all estates < 500 ha in each selected block. The Listing Survey involved traversing the block, largely on foot, and recording the name of each estate located, its owner's name, the address, the proportion of the estate falling within the block boundary, estate area and data on land tenure status. The position of each estate was roughly sketched on a map of the block and geographical co-ordinates of the estate obtained using GPS satellite receivers. The Listing Survey took approximately two months to complete. Estate numbers determined during this initial survey are shown in Table 2.

One concern in traversing a 10000 ha area to develop a sampling frame was its large dependence on the field staffs' motivation and interest in carrying out the work carefully and conscientiously. This was difficult to supervise, but prior training, a good allowance for field work and the requirement to bring back results which were difficult to falsify (particularly GPS coordinates of estate location) produced information that appeared to have a high level of accuracy.

An interesting feature that arose during the Listing Survey was the discovery of "ghost" estates that had been abandoned or left dormant by the owner. Decisions had to be made as to whether these should be included in the survey results. A subsequent small survey showed that abandonment or dormancy was largely due to lack of capital or due to a land dispute. Two thirds of the abandoned estates and half of the dormant estates were found encroached by smallholder farmers.

Table 2. Estate Numbers available from Listing Survey

Region	District	Estate Size Category				Total
		< 20	20 - <40	40 - <100	100 - <500	
North	Rumphi	153	28	11	3	195
	Mzimba 1	147	70	27	14	258
	Mzimba 2	59	36	10	4	109
Central	Kasungu 1	397	151	42	23	613
	Kasungu 2	404	158	39	19	620
	Dowa	226	34	4	0	264
	Lilongwe	888	128	15	7	1038
	Nkhotakota	83	65	44	7	199
South	Mangochi 1	12	18	22	19	71
	Mangochi 2	21	15	18	13	67
	Machinga 1	18	36	15	19	88
	Machinga 2	7	7	3	6	23
	Zomba	3	1	0	2	6
TOTAL		2418	747	250	136	3551

At the third and final stage of sampling, a number of estates were selected at random from those located during the Listing Survey. The number selected from each district was decided using standard sample size calculations. These suggested 14, 12, 20, 27 and 75 estates respectively from the selected districts for each area size category. However it was clear from Listing Survey results that the total numbers in some cases were smaller than the sample sizes recommended. An ad hoc procedure was developed to overcome this problem. The formula below was used to determine the number for inclusion. In this formula, n_{oc} refers to the number recommended for selection in size category c, while N_C refers to the total number of estates found in the c^{th} size category in the sampled blocks in a particular district.

$$\nu_c = \begin{cases} n_{oc} & \text{if} \quad N_c > 200 \\ \dfrac{n_{oc}}{1+\dfrac{n_{oc}}{N_c}} & \text{if} \quad 2\,n_{oc} < N_c \le 200 \\ \dfrac{2}{3}\,n_{oc} & \text{if} \quad (2/3)\,n_{oc} < N_c \le 2\,n_{oc} \\ N_c & \text{if} \quad N_c \le (2/3)\,n_{oc} \end{cases}$$

These problems did not arise in the case of sample size calculations of estates in the ≥ 500 ha category. The cleaned up database from the Ministry of Agriculture showed 173 such estates from

which 53 estates were selected at random from each region in proportion to the total number of estates (32, 103 and 38 in North, Central, South respectively) of size ≥ 500 ha in each region.

In every size category, not all estates selected provided data for the survey, largely owing to difficulties in locating the estate owner or manager. Where available, neighbouring estates of the same size category were then used. Out of 523 estates planned for inclusion in surveying estates in the < 500 ha size category, 519 were visited. In the case of estates in the ≥ 500 ha category, some estates were found to be of size < 500 ha. Replacement sites were then selected from the database and visited.

6. Success/inadequacy of sampling scheme and lessons learnt

Overall, the sampling procedures adopted appeared to be effective and produced sensible results with standard errors that were not unduly large. For example, regional and national level estimates for total land area covered by estates led to coefficients of variation (CVs) between 6% and 10%. For estimates of various types of land use (e.g. areas under annual crops, areas of under-utilised suitable land, areas under plantation forests), the standard errors were slightly higher, with CVs up to about 28%.

The Listing Survey proved to be an effective means of identifying the small estates (< 500 ha). It was invaluable in the estate selection process and served as a check against farmers' estimates of their land areas against more exact measurements using GPS equipment and aerial photos. On the other hand, the selection of estates of size > 500 ha, via database information held at the Ministry of Agriculture, was less effective. Firstly, the "cleaning" process applied to the database information proved to be time consuming and required discussions with many personnel with knowledge of the estate sector. Another difficulty occurred during the survey process when many of the estates visited were found to be smaller than 500 ha.

Retrospectively, we were also strongly criticised by several of the stakeholders attending the final ELUS workshop for conducting the survey of tea estates through use of a postal questionnaire. As a results, the ELUS project was extended by three months to address this concern and gather further information by visiting some of the tea estates.

There were also a number of questions raised at the final workshop concerning the sampling procedures. One question related to variation in the sampling fractions for estates of different size categories (see Table 3 below). It was felt that there was an under-representation of estates among the smaller size categories and over-representation among the larger size categories. We explained that differences in variability of estate areas in the different size categories led to different sampling fractions in order to produce results with the same level of precision. In a developing country context, it was clear that methodologies used had to be carefully described so that the reasons for adopting a particular methodology were clear to the non-statistically minded person.

Table 3. Estates listed in the Listing Survey and numbers actually visited.

	Estate Size Category				
	< 20	**20 - <40**	**40 - <100**	**100 - <500**	**Total**
Estates listed	2418	747	250	136	3551
Estates visited	158	120	128	114	519
Sampling fraction	0.0653	0.1606	0.512	0.838	0.146

Another lesson is the importance of not assuming that the existence of data means that it is available for use by external consultants. Obtaining permission to use even the minimum of data (at the planning stage, all we wanted were approximate figures for numbers of estates in each district) can be a time consuming process.

The time required to computerise and check a vast quantity of field data must also not be under-estimated. In the case of the ELUS survey, we found that after completion of data collection activities, and following computerisation of the results, another three to four months were needed to make quality checks and "clean" up the data. This often involved returning to the original records and consulting the field staff to clarify oddities.

7. Results generated by ELUS activities

Using the sampling protocol as planned, two parallel but separate surveys were conducted, namely a socio-economic survey and a land use survey. Of estates visited within each survey, 510 were common to both surveys. Further sub-sample surveys were conducted for more in-depth information, namely a Tenant and Labour Survey, a Nutrition and Food Security Survey, a Farm Management Survey, a Sugar Estates Survey, a Tea Estate Survey and a survey of abandoned estates. Each of these generated a substantial report of the findings and provided details to inform the Land Utilisation Steering Committee.

A data archive was also prepared, containing all information generated by ELUS activities. The estate identification codes were anonymised in the more public version of the archive. A training workshop was held to ensure that stakeholders from various government ministries and departments, with an interest in the data, were able to access the information with ease. Copies of the data archive were distributed to these stakeholders and other relevant organisations.

Acknowledgements:

Sincere thanks are due to the Land Use Specialist Steve Gossage for his very valuable help during work on the statistical aspects of ELUS and to Tom Barrett, DFID Natural Resources Adviser in Malawi in 1995, for his support to the project and for suggesting this topic as a case study of good biometric practices. Thanks are also due to the staff of the Estate Extension Services Trust, the Ministry of Agriculture and the Ministry of Lands and Valuation in Malawi for their assistance in various ways. The material in this paper however remains the sole responsibility of the author (an SSC staff member) and does not imply the expression of any opinion whatsoever on the part of the Government of Malawi or of the funding agency, the UK Government Department for International Development.

References:

ITAD (1993). Estate Extension Services Trust (EEST) – Skills Gap Survey. Report of Survey Results.

Rao, J.N.K. (1975). Unbiased variance estimation for multistage designs. *Sankhya* C, 37, 133-139.