Chapter V
Design of master sampling frames and master samples for household surveys in developing countries

Hans Pettersson
Statistics Sweden
Stockholm, Sweden

Abstract

The present chapter addresses issues concerning the design of master sampling frames and master samples. The introduction is followed by several sections. Section B gives a brief account of the reasons for developing and utilizing master sampling frames and master samples; section C contains a discussion of the main issues in the design of a master sampling frame; and section D covers master samples and addresses the important decisions to be taken during the design stage (choice of PSUs, number of sampling stages, stratification, allocation of sample over strata, etc.).

Key terms: master sampling frame, master sample, sample design, multistage sample.

-+

A. Introduction

1. National statistics offices (NSOs) in developing countries are usually the main providers of national, "official" statistics. In this role, the NSOs must consider a broad scope of information needs in the areas of demographic, social and economic statistics. The NSOs use different data sources and methods to collect the data. Administrative data and registers may be available to some extent but sample surveys will always be an important method of collection. Most NSOs in developing countries carry out several surveys every year. Some of the surveys (for example, the Living Standards Measurement Study, the Demographic and Health Survey, the Multiple Indicator Cluster Survey) are fairly standardized in design, while others are "tailor-made" to fit specific national demands. The need for planning and coordination of the survey operations has stimulated efforts to integrate the surveys in household survey programmes. Ad hoc scheduling of surveys has now been replaced in many NSOs by long-range plans in which surveys covering different topics are conducted continuously or at regular intervals. The United Nations National Household Survey Capability Programme (NHSCP) has played an important role in this process.

2. A household survey programme allows for integration of survey design and operations in several ways. The same concepts and definitions can be used for variables occurring in several surveys. Sharing of survey personnel and facilities among the surveys will secure effective use of staff and facilities. The integration may also include the use of common sampling frames and samples for all the surveys in the survey programme. The development of a master sampling frame (MSF) and a master sample (MS) for the surveys is often an important part of an integrated household survey programme.

3. The use of a common master sampling frame of area units for the first stage of sampling will improve the cost-efficiency of the surveys in a household survey programme. The cost of developing a good sampling frame is usually high; the establishment of a continuous survey programme makes it possible for the NSO to spread the costs of construction of a sampling frame over several surveys.

4. The cost-sharing can be taken a step further if the surveys select their samples as subsamples from a common master sample selected from the MSF. The use of a master sample for all or most of the surveys will reduce the costs of sample selection and preparation of sampling frames in the second and subsequent stages of selection for each survey. These cost advantages with the MSF and the MS also apply to unanticipated ad hoc surveys undertaken during the survey programme period and, indeed, also in the case where no formal survey programme exists at the NSO.

5. The present chapter will address issues concerning the design of master sampling frames and master samples for household surveys. The United Nations manual, *National Household Survey Capability Programme: sampling frames and sample designs for integrated household survey programmes* (United Nations, 1986) contains a good description of the various steps in the process of designing, preparing and maintaining a master sampling frame and a master sample. The manual includes an annex with several case studies. The interested reader is referred to that publication for a detailed treatment of the subject.

B. Master sampling frames and master samples: an overview

1. Master sampling frames

6. As described in chapter II, household samples in developing countries are normally selected in several sampling stages. The sampling units used at the first stage are called primary sampling units (PSUs). These units are area units. They can be administrative subdivisions like districts or wards or they can be areas demarcated for a specific purpose like census enumeration areas (EAs). The second stage consists of a sample of secondary sampling units (SSUs) selected within the selected PSUs. The last-stage sampling units in a multistage sample are called ultimate sampling units (USUs). A sampling frame - a list of units from which the sample is selected - is needed for each stage of selection in a multistage sample. The sampling frame for the first-stage units must cover the entire survey population exhaustively and without overlaps, but the second-stage sampling frames would be needed only within PSUs selected at the preceding stage.

7. If the PSUs are administrative units, a list of these units may exist or such a list could generally easily be assembled from administrative records for use as a sampling frame. Such an ad hoc list of PSUs could be prepared on every single occasion when a sample is needed. However, when there is to be a series of surveys over a period, it would be better to prepare and maintain a master sampling frame that is at hand for every occasion. The cost savings could be considerable compared with ad hoc preparation of sampling frames for each occasion. Also, the fact that the frame will be used for a number of surveys will make it easier to justify the costs of its development and maintenance and to motivate spending resources on improvements of the quality of the frame.

8. A master sampling frame is basically a list of area units that covers the whole country. For each unit there may be information on urban/rural classification, identification of higher-level units (for example, the district and province to which the unit belongs), population counts and, possibly, other characteristics. For each area unit, there must also be information on the boundaries of the unit. The MSF for the household surveys in the Lao People's Democratic Republic, for example, contains a list of approximately 11,000 villages. For each village, there is information on the number of households, number of females and males, whether the village is urban or rural (administrative subdivisions in urban areas are also called villages) and information on which district and province the village belongs to. There is also information on whether the village is accessible by road.

9. The most common type of MSF is one with EAs as the basic frame units. Usually, there is information for each unit that links the unit to higher-level units (administrative subdivisions). From such an MSF, it is possible to select samples of EAs directly. It is also possible to select samples of administrative subdivisions and to select samples of EAs within the selected subdivisions.

10. An up-to-date MSF with built-in flexibility has advantages apart from the cost and quality aspects discussed above. It facilitates quick and easy selection of samples for surveys of different kinds and it could meet different requirements for the sample from the surveys. Another

advantage is that a well-maintained MSF will be of value for the next population census. The census itself requires a frame similar to the frame that will be used for household surveys. The job of developing the frame for the census is likely to be considerably easier if a well-kept master sampling frame has been in use during the intercensal period. The ideal situation is one where the new MSF is planned and constructed during the census period and then fully updated during the next census.

2. Master samples

11. From a master sampling frame, it is possible to select the samples for different surveys entirely independently. However, in many instances, there are substantial benefits resulting from selecting one large sample, a master sample, and then selecting subsamples of this master sample to service different (but related) surveys. Many NSOs have decided to develop a master sample to serve the needs of their household surveys.

12. A master sample is a sample from which subsamples can be selected to serve the needs of more than one survey or survey round (United Nations, 1986), and it can take several forms. A master sample with simple and rather common design is one consisting of PSUs, where the PSUs are EAs. The sample is used for two-stage sample selection, in which the second-stage sampling units (SSUs) are housing units or households.

13. The subsampling can be carried out in many different ways. Subsampling on the primary level (of PSUs) would give a unique subsample of the master sample PSUs for each survey, that is to say, each survey would have a different sample of EAs. Subsampling on the secondary level would give a subsample of housing units from each master sample PSU, that is to say, each survey would have the same sample of EAs but different samples of housing units within the EAs. The subsampling could be carried out independently, or some kind of controlled selection process could be employed to ensure that the overlap between samples will be on the desired level. Another way of selecting samples from the master sample would be to select independent replicates from the sample. One or several of the replicates could be selected as a subsample for each survey. Such a set-up would require that the master sample be built up from the start from a set of fully independent replicates.

14. An NSO can reap substantial cost benefits from the use of a master sample. The costs of selecting the master sample units will be shared by all the surveys using the MS; the sample selection costs per survey will thus be reduced. Since the selection of master sample units is basically an office operation (especially if a good MSF exists), the cost savings at this stage may be modest. Much greater cost savings are realized when the costs for preparing maps and subsampling frames of housing units within master sample units are shared by the surveys. The fieldwork required to establish subsampling frames is usually extensive; and the cost per survey of this fieldwork will decrease almost proportionally to the number of surveys using the same subsample frame.

15. In some countries, the difficulties and the costs related to travel in the field might make it economical to recruit interviewers within or close to the MS primary sampling units and have them stationed there for the whole survey period. In that case, relatively large PSUs are used.

There is then a clear gain to be derived from using a fixed master sample of such PSUs rather than selecting a new sample for each survey and having to relocate the interviewers or recruit new interviewers.

16. The use of the same master sample units will reduce the time it takes to get the surveys started in the area. In many developing countries, the interviewer needs to secure permission from regional and local authorities to conduct the interviews in the area. In countries like the Lao People's Democratic Republic and Viet Nam, for example, permits need to be obtained at several administrative levels down to that of the village chairman. The time required for this process of "setting up shop" will be reduced substantially when the same areas are used for several surveys.

17. The use of the same master sample PSUs for several surveys will reduce the time that it takes for the interviewer to find the households. When maps and subsampling frames of good quality are available, the interviewer can quickly navigate the area; in some cases, he or she may even have worked in the area during a previous survey. A permanent numbering of housing units may be introduced to facilitate orientation in the area. This has been done in some master samples: Torene and Torene (1987) describe the case of the Bangladesh master sample.

18. The MS makes it possible to have overlapping samples in two or more surveys. This permits integration of data at the microlevel through the linking of household data from the surveys. There is a risk, however, of adverse effects on the quality of survey results when sample units are used several times. Households participating in several rounds of a survey or in several surveys may become reluctant to participate or may be less inclined to give accurate responses in the later surveys.

19. An MS thus has advantages (costs, integration and coordination) for the regular surveys in a survey programme. An MS that is in place will also allow the NSO to be better prepared to handle sampling for ad hoc surveys: subsamples can be selected quickly from the MS when they are needed for ad hoc surveys.

20. The advantages of master samples are apparent but there are also some disadvantages or limitations. The master sample design always represents a compromise among different design requirements arising from the surveys in the programme. The master sample will suit surveys that have reasonably compatible design requirements with respect to domain estimates and the distribution of the target population within those areas. The design chosen for the master sample will usually suit most of the surveys in the survey programme fairly well, but none perfectly. The master sample design imposes constraints and requirements (concerning sample size, clustering, stratification, etc.) on the individual surveys that sometimes can be difficult to accommodate. This will result in some loss of efficiency in the individual surveys.

21. There are also surveys with special design requirements that the master sample will not be able to accommodate at all, namely:

- Surveys aimed at certain regional or local areas where a large sample is needed for a small area (for example, surveys used for assessing the effects of a development project in a local area).

- Surveys aimed at unevenly distributed population (for example, ethnic) subgroups.

22. An example of the first type is the survey of opium-growing that is conducted regularly in some areas in four northern provinces in the Lao People's Democratic Republic. The purpose is to evaluate the progress of the Lao government project aiming at reducing opium-growing. In this case, since the Lao master sample could not meet the demands on the sample design, a separate sample was selected for the survey. (An alternative would have been to use the master sample PSUs in the four provinces and to select additional PSUs from the master sample frame.)

23. In some cases, the cost savings of a master sample may not be realized fully. To draw a subsample from a master sample to suit the specific needs of an individual survey and then to compute the selection probabilities correctly require technical skills. This can be a more complicated operation than selecting an independent sample. The fact that sampling statisticians are scarce in many NSOs in developing countries may hamper the use of a master sample or, indeed, hinder the development of a master sample. There are examples of master samples that are underutilized owing to the lack of sampling competence at the NSO.

3. Summary and conclusion

24. The advantages, disadvantages and limitations discussed above can be summarized as follows:

Master sampling frame:

- Cost efficient; makes it possible for the NSO to spread the costs of construction of a sampling frame over several surveys.

- Quality will usually be better than that of ad hoc sampling frames because it is easier to motivate investments in quality improvement in a frame that will be used over a longer period.

- Simplifies the technical process of drawing individual samples; facilitates quick and easy selection of samples for surveys of different kinds.

- If well-maintained, it will be of value for the next population census.

Master sample:

- *Cost savings:*
  - Costs of selecting the master sample units will be shared by all the surveys using the MS.
  - Costs of preparing maps and subsampling frames of dwelling units or households will be shared among the surveys using the MS; however, subsampling frames will need to be updated periodically to add new construction and remove demolished housing units.
    - Clear gain from using an MS in the case where interviewers need to be stationed in or close to the PSU owing to difficulties and high costs related to travel in the field.
- *More efficient operations:*
  - Use of the same master sample PSUs for several surveys will reduce the time it takes to get the surveys started in the area and also the time it takes the interviewer to find the respondents.
  - The MS facilitates quick and easy selection of samples; subsamples from the MS can be selected quickly when needed for ad hoc surveys.
- *Integration:*
  - That the MS makes it possible to have overlapping samples in two or more surveys, provides for integration of data from the surveys.
- *Limitations, disadvantages:*
  - The MS will not be suitable for all surveys; in some cases, the NSO will face situations during the survey programme period where unanticipated survey needs arise that cannot be met by a master sample (this is a limitation and not really a disadvantage).
  - When sample units are reused, especially at the household level, there are risks of biases resulting from conditioning effects and from increased non-response caused by the cumulative response burden.
  - The continuous operation of an MS requires sampling skills that may not be available at the NSO.

Conclusion

25. It is apparent that master sampling frames and master samples have many attractive features. It is desirable for every NSO to have a well-kept master sampling frame that can cater for the needs of its household surveys, regardless of whether the surveys are organized in a survey programme or conducted in an ad hoc manner. Many NSOs will find it beneficial to take the further step of designing and using a master sample for all or most of the household surveys.

C. Design of a master sampling frame

26. The national household survey programme defines the demands on the master sampling frame and the master sample design in terms of, for example, the anticipated number of samples, population coverage, stratification and sample sizes. How these demands should be met in the design work depends on the conditions for frame construction in the country. The most important factor is the availability of data and other material that can be used for frame construction. In section 1 below, we discuss briefly the types of data and materials that are needed and the quality problems that may be present in the data.

27. When the available data and materials have been assessed, the NSO has to decide on the key characteristics of the MSF related to:

- Coverage of the MSF (see sect. 2)
- Which area units should serve as frame units in the MSF (see sect. 3)
- What information about the frame units should be included in the MSF (see sect. 4)

28. Complete, well-handled documentation of the frame, as well as clear procedures for updating, is crucial for efficient use of the MSF (see sect. 5).

1. Data and materials: assessment of quality

29. The most important source of data and materials will usually be the latest population census. This is obvious in the case where the NSO intends to use census enumeration areas as frame units; but even if other (administrative) units will be used, there is usually a need for population or household data from the census for them. The basic materials from the census are lists of EAs with population and household counts and sketch maps of the EAs. There are also maps of larger areas (districts, regions) on which the EAs are marked. Usually EAs are identified by a code showing urban/rural classification and the administrative division and subdivision to which they belong. Sometimes the code also shows whether the EA contains institutional population (living in military barracks, student hostels, etc.).

30. The quality of the census data and materials varies considerably from one country to another. This is especially true for the maps. Some countries, like South Africa, have digitized EA maps stored in databases while others, like the Lao People's Democratic Republic, have no

good maps at all. In some countries, the EA maps are often very sketchy and difficult to use in the field. As the EAs may actually be composed of lists of localities rather than of proper aeral units, scattered populations outside the listed localities may not be covered in such frames. A special quality-related problem that is somewhat annoying for the frame developer is difficulty in retrieving census materials, especially maps. The maps may be of good quality but this does not help if they are difficult to retrieve. The fact that it is still rather common for EA maps to be "buried" in an archive after the census, sometimes in less than good order, makes them difficult to find. It is also not uncommon for some EA maps to be missing from the archive.

31. Generally, the quality of the census material deteriorates over time. This is definitely the case with the population counts for EAs where population growth and migration will affect EAs unequally. Also, changes in administrative units, like boundary changes or splitting/merging of units, will cause the census information to become outdated. The census information is bound to be outdated if the last census was conducted seven or eight years before.

32. A first step in the design of the MSF must be to identify and assess the different materials available for frame construction, including not only the census materials but also other data/materials: even if the population census is to be the main source for materials, there are other sources that may be needed for updating or supplementing the census data. The questions to be asked are: What data/materials are available and how accurate are they?; and How current are the data and how often are they updated? Maps need to be evaluated regarding their amount of detail and to what extent the boundaries of administrative subdivisions are shown. Efforts should be made to estimate the proportion of EA sketch maps that meet required standards of quality.

33. At this stage of the work, it is also important to obtain or prepare a precise and thorough description of the administrative structure of the country and an up-to-date list of its administrative divisions and subdivisions.

2. Decision on the coverage of the master sampling frame

34. An early decision to be made concerns the coverage of the MSF. Should certain very remote and sparsely populated parts be excluded from the frame? The decision of most countries to have full national coverage in the MSF is generally a wise one because when certain remote and sparsely populated parts are excluded from the regular surveys in the programme, there may still arise situations where an ad hoc survey needs to cover these parts. A special case involves nomadic groups and hill tribes that are difficult to sample and to reach in the fieldwork. Such groups are excluded from the target population of the household survey programmes in some countries.

35. A decision must also be taken on the coverage of the institutional population. In some countries, large institutions are defined as special enumeration areas (boarding schools, large hospitals, military barracks, and hostels for mine workers). In that case, it would be possible to exclude these areas from the frame. In general, however, it is better to keep these units in the frame, thus providing room for coverage decisions in future surveys.

3. Decision on basic frame units

36. Frame units are the sampling units included in the master sampling frame. Basic frame units are the lowest-level units in the master sampling frame. Generally, it is desirable for the basic frame units to be small areas that will allow for a grouping of the units into larger sampling units if a certain survey's cost considerations should require this.

37. Census enumeration areas are often the best choice for basic frame units. The EAs have several advantages as basic frame units. The demarcation of EAs is carried out with the aim of producing approximately equal-sized areas in terms of population, which are an advantage in some sampling situations. The EAs are mapped; usually the map is supplemented by a description of the boundaries. Base maps showing the location of EAs within administrative divisions are usually available. Computerized lists of EAs are produced in the census; these lists can be used as the starting point for a MSF. There is much that weighs in favour of using EAs as frame units but quality problems of the kinds discussed in section 1 may in some cases lead to other solutions.

38. Some countries have administrative subdivisions that are small enough to serve as basic frame units; and there may be situations where these units have advantages over EAs as basic frame units, like that involving the MSF maintained by the National Statistics Centre in the Lao People's Democratic Republic. EAs had been considered basic frame units but it was found that the documentation of the EAs was difficult to retrieve, and generally of rather poor quality, making the EA boundaries difficult to trace in the field. In this situation, it was decided to use villages as basic frame units. The villages in the Leo People's Democratic Republic are well-defined administrative units. They are not, however, area units in a strict sense. The boundaries between villages are fuzzy and no proper maps exist, but there is no uncertainty about which households belong to a given village.

39. Cases where units smaller than EAs serve as basic frame units are not common but such cases do exist. An example is Thailand where the EAs in municipal areas are subdivided into blocks and census enumeration of population and households is carried out for each block. Those blocks were used as basic frame units in the municipal part of the MSF.

40. The basic frame units, whether EAs or other type of units, will differ in size in terms of number of households and population in the area. Even if the intention is to create EAs that do not show too much population-wise variation in size, there will be deviations from this rule for various reasons (for example, smaller EAs in terms of population may be constructed in sparsely populated areas where travel is difficult). The result is usually a substantial variation in EA size with some extreme cases at the low and high ends. In Viet Nam, for example, the average number of households per enumeration area is 100. The number of households in the 166,000 EAs varies from a minimum of 2 to a maximum of 304 (Glewwe and Yansaneh, 2001). Approximately 1 per cent of the EAs have 50 or fewer households. In the Lao People's Democratic Republic, the proportion of small EAs is even larger: 6 per cent of the EAs have less than 25 households. Such population-wise variation in the size of the areas that are used as basic frame units will generally not be a problem, but very small units are not suitable for use as

sampling units. Very small EAs can be accepted in the MSF; but for samples based on the MSF, these EAs need to be linked to adjacent EAs to form suitable sampling units.

4. Information about the frame units to be included in the frame

41. A simple list of the basic frame units constitutes a rudimentary sampling frame but the possibility of drawing efficient samples from such a frame is limited. The usefulness of the frame will be greatly improved if it contains supplemental data about the frame units that could be used to develop efficient sample designs. The supplemental data may be of three types:

(a) Information that makes it possible to group basic frame units into larger units. One way to increase the potential for efficient sampling from the frame is to allow sampling of different types of units from the frame. It is therefore desirable that the frame contain information that makes it possible to form larger units and thus achieve flexibility in the choice of sampling units from the frame;

(b) Information on size of the units. The efficiency of samples from the frame will also be enhanced if a measure of size is included for each frame unit. This is especially important when there is large variation in the sizes of the units*;*

(c) Other supplemental information. Information that could be used for stratification of the units or as auxiliary variables at the estimation stage will improve the efficiency of samples from the MSF.

Information that makes it possible to group basic frame units into larger units

42. For some surveys, the best alternative for PSUs is small areas like enumeration areas. For other surveys, considerations of costs and sampling errors will weigh in favour of PSUs that are considerably larger than EAs. These larger PSUs could be built from groups of neighbouring EAs. Another possibility is to use administrative units like wards and districts as PSUs. In all such cases, it is necessary that the master sampling frame provide possibilities for the construction of these larger PSUs. It is therefore important that the frame unit records in the MSF contain information on the higher-level units to which the frame unit belongs.

43. A model design of a master sampling frame that has been used by many countries is one that uses census enumeration areas as basic frame units and where the units are ordered geographically into larger (administrative) units in a hierarchic structure. Samples can be drawn from the MSF in different ways: (a) by sampling EAs; (b) by grouping EAs to form PSUs of convenient size and sampling the PSUs; and (c) by sampling administrative subdivisions at the first stage and subsequent sampling in additional stages down to the EA level. The hierarchic structure in the master sampling frame of Viet Nam contains the following levels:

Provinces
Districts
Communes (rural), wards (urban)
Villages (rural), blocks (urban)
Census enumeration areas

44. Flexibility in the choice of sampling units is further enhanced if all frame units (basic frame units as well as higher-level units) are assigned identifiers based on geographical adjacency. This makes it possible to use the frame units as building blocks to form PSUs of required size from adjacent frame units. Such an operation would be needed in the cases of Viet Nam and the Lao People's Democratic Republic described in the previous section. Another advantage with an identifier based on geographical adjacency is that geographically dispersed samples can be selected from the master sampling frame by the use of systematic sampling from geographically ordered sampling units.

Measures of size of frame units

45. The inclusion of measures of size is especially important if there is large variation in the size of the frame units. Usually, the measures of size are counts of population, households or dwelling units within the frame unit. It is important to note that measures of size do not need to be exact. In fact, they are virtually always inaccurate to some extent because they are based on data from a previous point in time and the fact that the population is ever-changing will gradually result in their becoming out of date. Errors in the measures of size do not lead to biases in the survey estimates but they do reduce the efficiency of the use of the measures of size, especially in the case where the measures of size are used at the estimation stage. Efforts should therefore be made to ensure that the measures of size are as accurate as possible.

46. Measures of size are most commonly used in the sample selection of frame units with probability proportional to size (PPS). Other uses of measures of size are:

- To determine the allocation of sample PSUs to strata
- To form strata of units classified by size
- As auxiliary variables for ratio or regression estimates
- To form sampling units of a desirable size

Other supplemental data for the frame units

47. Supplemental information about the frame units that could be obtained at reasonable costs should be considered for inclusion in the frame. Information on population density, predominant ethnic groups, main economic activity and average income level in the frame units are variables that are often useful for stratification.

48. In the Namibia master sampling frame, a crude income-level classification into high income, medium income, and low income was included for the urban basic frame units (EAs) in the capital, Windhoek, making it possible to form two income-level strata in the urban sub-domain of Windhoek. Another example is the Lao master sampling frame where the rural frame units have information on whether the unit is close to a road or not. The samples for the household surveys using the master sampling frame are stratified on access/no access to a road.

5. Documentation and maintenance of a master sampling frame

Documentation

49. A well-kept, accurate and easily accessible documentation of the master sampling frame is imperative for the use of the frame. If the documentation is poor, the benefits of the frame will not be fully realized. The core of the documentation is a database containing all the frame units. The contents of the records for frame units should be:

- A primary identifier, which should be numerical. It should have a code that uniquely identifies all the administrative divisions and subdivisions in which the frame unit is located. It will be an advantage if the frame units are numbered in geographical order. Usually EA codes have these properties. Fully numerical identifiers are better than names or alphanumeric codes. In many cases, existing geo-coding systems from administrative sources and from the census will be suitable as primary identifiers.

- A secondary identifier, which will be the name of the village (or other administrative subdivision) where the frame unit is located. Secondary identifiers are used to locate the frame unit on maps and in the field.

- A number of unit characteristics, such as measure of size (population, households), urban/rural, population density, etc. All data concerning the unit that could be obtained at a reasonable cost and having acceptable quality should be included. The characteristics could be used for stratification, assigning selection probabilities, and as auxiliary variables in the estimation.

- Operational data, information on changes in units and indication of sample usage.

50. The frame must be easy to access and to use for various manipulations like sorting, filtering and production of summary statistics that can help in sample design and estimation. That is best done if the frame is stored in a computer database. The use of formats that can be accessed only by specialists should be avoided. A simple spreadsheet in Excel will often serve well. Excel is easy to use, many know how to use it, and it has functions for sorting, filtering and aggregation that are needed when samples are prepared from the frame. The worksheets could easily be imported in most other software packages.

Maintaining the MSF

51. Closely linked to the documentation of the MSF are the routines for maintaining the frame. During the time of use of the MSF, changes will occur that affect both the number and the definition of the frame units. The amount of work required to maintain a master sampling frame depends primarily on the stability of the frame units. There are two kinds of changes that may occur in the frame units: changes in frame unit boundaries and changes in frame unit characteristics.

52. Frame unit boundary changes affect primarily administrative subdivisions. Administrative subdivisions are subject to boundary changes, especially at the lower levels, owing to political or administrative decisions. Often these changes are made in response to substantial changes of the population of the areas affected. New units are created by splitting/combining existing units or by more complicated rearrangements of the units. Also, boundaries of existing units may be altered without creation of any new units. If there are frequent changes in administrative subdivisions, considerable resources have to be allocated to keep the frame up to date and accurate.

53. Changes affecting the boundaries of frame units must be recorded in the MSF. A system for collecting information about administrative changes needs to be established to keep track of these changes.

54. Changes in frame unit characteristics include not only simple changes such as name changes but also more substantial changes like changes in the measure of size (population or number of households/dwelling units) or changes in urban/rural classification. These changes do not necessarily have to be reflected in the MSF. However, as has been said above, outdated information on measures of size results in a loss of efficiency in the samples selected from the frame. Updating measures of size for the whole frame would be very costly and generally not cost-efficient; but for especially fast-growing peri-urban areas, it is a good idea to update the measures of size regularly.

55. Changes in measures of size for frame units become problematic when there are large and sudden changes in the population, which may occur, for example, in squatter areas when local authorities decide to remove the squatters from the area. Such dramatic changes need to be reflected in the sampling frame. An example of a less dramatic but still problematic change (for the sampling frame) is the Government-initiated migration from remote villages in the mountainous areas of the Lao People's Democratic Republic. The Government is encouraging the members of these villages to move to villages with better access to basic services. As a result of this process, the number of villages has declined by approximately 10 per cent over a two-year period. Clearly these changes must be included in the sampling frame.

56. There is a risk that the maintenance of the MSF will be neglected when a NSO is operating with scarce resources and is struggling to keep up with the demand for statistical results. It is therefore important that the NSO develop plans and procedures for frame updating at an early stage and that sufficient resources are allocated for the purpose.

D. Design of master samples

57. A master sample is a sample from which subsamples can be selected to serve the needs of more than one survey or survey round (United Nations, 1986). The main objective should be to provide samples for household surveys that have reasonably compatible design requirements with respect to domains of analysis and the distributions of their target populations within those areas. The master sample is defined in terms of the number of sampling stages and the type of units that serve as ultimate sampling units (USU). A master sample selected in two stages with enumeration areas as the second stage units would be called a *two-stage master sample of enumeration areas.* If the EAs were selected directly at the first stage, we would have a *one-stage master sample of EAs*. Both these designs are common master sample designs in developing countries.

58. Important steps in the development of a master sample are discussed in sections D.1-D.4. In sections D.5 and D.6, issues concerning the documentation and maintenance of the master sample are discussed. Finally, section D.7 discusses the use of the master sample for surveys that are not primarily aimed at households.

1. Choice of primary sampling units for the master sample

59. The MSF provides the frame for the selection of the master sample. The basic frame unit in the MSF could, in some cases, be used as the primary sampling unit for the master sample. In other cases, we may decide to form PSUs that are larger than the basic frame units in the MSF. In these cases, usually some kind of well-defined administrative units (counties, wards, etc.) are used as PSUs; but there are also cases where the PSUs have been constructed by using the frame units as building blocks. In this case, adjacent units are grouped into PSUs of convenient size. One example is the Lesotho master sample where the PSUs were formed by combining adjacent census EAs into groups consisting of 300-400 households. The 3,055 census EAs were grouped into 1,038 EA groups which were to serve as PSUs (Pettersson, 2001).

60. There are several factors relating to statistical efficiency, costs and operational procedures to be taken into account when deciding on what should be the primary sampling unit. Assuming that the basic frame units in the MSF are EAs, under what circumstances would we prefer to use units larger than EAs as PSUs?

- If we know that the demarcations of a significant proportion of EAs are of poor quality, we may decide to use larger units as PSUs since larger areas generally provide more stable and clearly demarcated boundaries.

- When travel between areas is difficult and/or expensive. The difficulties and the costs related to travel in the field might make it economical to recruit interviewers within or close to the sampled PSUs and have them stationed there for the whole survey period. This would call for rather large PSUs.

- When the usage of the PSU for samples will be so extensive that a small PSU like an EA will quickly become exhausted. This problem could be solved either by using larger units as PSUs or by keeping the EAs as PSUs and rotating the sample of EAs. The first option is preferable when the cost of entering and launching the survey in the area is high.

- When, for reasons of cost control and sampling efficiency, it is customary to introduce one or more sampling stages involving units that are larger than the basic frame units. If, for example, the basic frame units are EAs, we may decide to use larger units, for example, wards, as PSUs and then select EAs or other area units within PSUs in the next stage.

- When, as in some surveys, household and individual variables are linked to community variables. An example is a health survey where individual health variables are linked to variables concerning health facilities in the village or commune. Another example is a living standards survey where household variables are linked to community variables on schools, roads, water, sanitation, local prices, etc. If the master sample should serve several surveys of this kind, there are advantages in using the community (village, commune, ward etc.) as the PSU. If the community is used as PSU, we can make sure that the subsample of SSUs will be well spread over the community.

61. Large area units are not suitable as PSUs because there are too few of them. It would not be meaningful to sample from a population of 50-100 units. Preferably, the number of PSUs in the population should be over 1,000 so that a 10 per cent sample will yield over 100 PSUs for the sample. A much larger fraction than 10 per cent would reduce the cost benefits of sampling. A much smaller number of PSUs than 100 in the sample would increase the variance. It should also be pointed out that it could be efficient to use different types of PSUs in different parts of the population, for example, EAs in urban areas and larger units in rural areas.

2. Combining/splitting areas to reduce variation in PSU sizes

62. When a decision has been reached concerning which type of unit should serve as PSU (and, in the case of two area stages, which unit should serve as SSU), we may find that there are "outliers" that are much smaller or larger than what is desirable.

Very small sampling units

63. Very small PSUs in the master sample are problematic. What should be considered acceptable size depends on the intended workload for the master sample. Statistics South Africa, which is using census EAs as PSUs for its master sample, decided to have 100 households as the minimum size of the PSUs. EAs having less than 100 households were linked with neighboring EAs during the preparation of the MSF. For its master sample, the National Central Statistics Office of Namibia applied the rule that the PSUs should contain at least 80 households. In the census, 2,162 EAs were formed. After joining the small EAs to adjacent ones, 1,696 PSUs

remained. Of the 1,696 PSUs, 405 were formed by joining several EAs; each of the remaining 1,291 consisted of a single EA.

64. The job of linking small EAs before selection can be very demanding if the number of small EAs is large. The case of Viet Nam can be taken as an example. For its surveys, the General Statistical Office of Viet Nam wanted a sample of areas with at least 70-75 households. Approximately 5 per cent of the EAs (= 8,000 EAs) have less than 70 households (Pettersson, 2001). The job of combining approximately 8,000 EAs with adjacent EAs was a tedious and time-consuming task.

65. One way to reduce the work of combining the small area units into fair-sized PSUs is to carry out this operation only when a small area (PSU) happens to be selected into the sample. Kish (1965) designed a procedure for linking small PSUs with neighbouring PSUs during or after the selection process.

66. Another way to reduce the work of combining small units is to introduce a sampling stage above the intended first stage. Instead of using the intended area units as PSUs, we could, in some cases, use larger areas as PSUs. In the selected PSUs, we carry out the operation of combining small area units (our originally intended PSUs) into fair-sized area units. The work of combining small area units is done only within the selected first-stage units, thus reducing the work considerably in this case, compared with the situation where we use the smaller areas as first-stage units. This alternative involves an additional sample stage above the intended first stage, which may affect the efficiency of the design. However, if we select only one SSU per selected PSU at the second stage, the sample will in effect be equivalent to the intended one-stage sample of area units. This was the solution used in the Vietnamese case. It was decided to use larger administrative units, namely, communes, instead of EAs, as the PSUs. Within the selected communes, the undersized EAs were linked to adjacent EAs to form units of acceptable size. In this way, the work of linking small EAs to adjacent EAs was reduced. Instead of linking 8,000 EAs, the work was confined to linking approximately 1,400 EAs in 1,800 selected communes. Three EAs (or EA groups in the case of small EAs) were selected at the second stage in the selected communes.

Very large area units

67. At the other extreme, there may be cases of area units that are too large -- in terms either of population or of geographical area -- to serve as PSUs. In both cases, the listing costs will be much greater than for the ordinary area units (EAs or some other area units). Problems will arise in both cases if some of the very large PSUs are selected for the master sample. In order to reduce the work of preparing list frames of households in these large units, we can put the large units in separate strata and select these PSUs with reduced sampling rates; we could maintain the overall sampling rates by increasing the sampling rates within PSUs.

68. Another way of handling the problem with a large PSU is to divide the PSU into a number of segments and select one segment randomly. The problem is a bit simpler than the problem with small PSUs, mainly because we do not have to take any action prior to the

selection of the master sample. Only when we happen to select a large PSU for the master sample do we need to take action.

69. A separate problem concerns PSUs that have grown or declined markedly since the time of the census. There will always be changes in population over time making the PSU measures of size less accurate over time. The general effect is an increase in variances; however, no bias is introduced. The problem becomes a serious one when dramatic changes occur in some PSUs owing, for example, to clearing of suburban areas or large-scale new construction in some areas. Procedures for handling these changes have to be designed as a part of the maintenance of the master sample. The NHSCP manual discusses two strategies: sample replacement and sample revision (United Nations, 1986).

3. Stratification of PSUs and allocation of the master sample to strata

Stratification

70. The master sample PSUs are often stratified into the main administrative divisions of the country (provinces, regions, etc.) and within these divisions, into urban and rural parts. Other common stratification factors are urbanization level (metropolitan, cities, towns, villages) and socio-economic and ecological characteristics. In the Lesotho master sample, the PSUs are stratified on 10 administrative regions and 4 agro-economic zones (lowland, foothill, mountain, and Senqu River valley), resulting in 23 strata that reflect the different modes of living in the rural areas.

71. It is possible to define "urban fringe" strata in rural areas close to large cities. This will take care of rural households that are, to some extent, dependent on the modern sector. In large cities, a secondary stratification could be carried out according to housing standard, income level or some other socio-economic characteristics.

72. A common technique used to achieve a deeper stratification within main strata is to order the PSUs within strata according to a stratification criterion and to select the sample systematically (implicit stratification). One advantage with implicit stratification is that the boundaries of the strata do not need to be defined.

Sample allocation

73. The allocation of master sample PSUs to strata could take different forms:

- Allocation proportional to the population in the strata
- Equal allocation to strata
- Allocation proportional to the square root of the population in the strata

74. Many master samples are allocated to the strata proportionally to the population (number of persons or households) in the strata. Proportional allocation is a sound strategy in many

situations. However, the proportional allocation assigns a small proportion of the sample to small strata. This may be a problem when the main strata are administrative regions (for example, provinces) of the country for which separate survey estimates are required and when the sizes of these regions differ greatly in size (as is often the case). The demand for equal allocation of the sample across provinces could be very strong among top government officials in the provinces (at least officials in the small provinces). When the provinces differ greatly in size, the equal allocation will result in substantial variation in sampling fractions between provinces. In the Lao master sample constructed in 1997, it was decided to use equal allocation across the 19 provincial strata in order to achieve equal precision for the province estimates. This resulted in sampling fractions where the smallest province had a sampling fraction 10 times larger than the fraction for the most populous province.

75. A strict proportional allocation over urban/rural domains will result in small urban samples in countries with small urban populations. The master sample prepared by the National Institute of Statistics of Cambodia is allocated proportionally over provinces and urban/rural. The sample of 600 PSUs consists of 512 rural and 88 urban PSUs. For some surveys, the urban sample has been considered too small and additional sampling of urban PSUs has been required. It may have been wise to oversample the urban domain somewhat in the master sample.

76. A compromise between the proportional and the equal allocation is the *square root allocation* where the sample is allocated proportionally to the square root of the stratum size. Square root allocation has been used for the master samples in Viet Nam and South Africa. Kish (1988) has proposed an alternative compromise based on an allocation proportional to $n\sqrt{(W_h^2 + H^{-2})}$ where n is the overall sample size, W_h is the relative size of stratum h and H is the number of strata. For very small strata, the second term dominates the first, thereby ensuring that allocations to the small strata are not too small.

77. Another compromise would be to have a large master sample suitable for province-level estimates and a subsample from the large sample that would mainly be designed for national estimates. An example is the 1996 master sample of the Philippines which consisted of 3,416 PSUs in an expanded sample for provincial-level estimates with a subsample of 2,247 PSUs designated as the core master sample in cases where only regional-level estimates were needed.

4. Sampling of PSUs

78. The most common method is to select the master sample PSUs with probability proportional to size (PPS). In this case, the probability of selecting a PSU is proportional to the population of the PSU, giving a large PSU a higher probability of being included in the sample.

79. The method has some practical advantages when the PSUs vary considerably in size. First, it could lead to self-weighting samples. Second, it generates approximately equal sample sizes within PSUs, which in turn implies approximately equal interviewer workloads, a desirable situation from a fieldwork perspective. More details on PPS sampling and its advantages and limitations are provided in chapter II.

80. A PPS sample can be selected in a number of ways. A common method is systematic selection within strata. If the PSUs are listed in some kind of geographical order within strata, this would result in a good geographical spread of the sample within the main strata (more details are provided in chap. II). The master samples of Lesotho, the Lao People's Democratic Republic and Viet Nam are all selected with systematic PPS with one random starting point within each stratum.

Interpenetrating subsamples

81. An alternative means of selecting the sample entails selecting a set of interpenetrating subsamples. An interpenetrating subsample is one subsample of a set of subsamples each of which constitutes, by itself, a probability sample of the target population.

82. The possibility of using interpenetrating subsamples when subsampling the master sample has some advantages. The subsamples provide flexibility in sample size. The sample for a particular survey can be made up of one or several of the subsamples. The subsamples can also be used for sample replacement in multi-round surveys.

83. The use of interpenetrating subsamples in the master sample design is not as common as the use of simple systematic selection. One example of a master sample using interpenetrating samples is that developed by the Statistics Office of Nigeria (Ajayi, 2000).

5. Durability of master samples

84. The quality of the master sample deteriorates over time; but the fact that the measures of size used for assigning selection probabilities become out of date as population changes take place would not be a problem if the population change were a more or less uniform growth in all units in the master sampling frame. However, this is usually not the case. Population growth and migration occur at varying rates in different areas: often there is low growth, or even a decline, in some rural areas, and high growth in some suburban areas in the cities. When such uneven growth takes place, the measures of size used in the selection of the master sample will cease to reflect the relative distribution of the survey population. This leads to increased sampling errors of estimates from the master sample. Also, changes in administrative boundaries and classifications (for example urban/rural classification of areas) may cause the stratification to become out of date.

85. The master sampling frame is normally completely revised after each population census, usually every 10 years. During the intercensal period, the frame should be updated regularly. The availability of a well-kept, regularly updated master sampling frame makes it possible to select entirely new master samples periodically from the master sampling frame. The question then is, For how long should a master sample be kept without significant changes? The durability of a master sample depends, to some extent, on local conditions such as internal migration and the rate of changes in administrative units. It is thus not possible to give a general recommendation that fits all situations. Often, the efficiency of a master sample will have deteriorated

substantially after three to four years. The decision to use the master sample without adjustments for a longer period needs to be carefully considered.

86. There are basically two strategies for handling the problem of deteriorating efficiency in the master sample. One is to select an entirely new master sample at regular intervals; in Lesotho, for example, the master sample is replaced every third year. The other strategy is to retain the master sample for a longer period but to make regular adjustments to compensate for the effects of changes in the frame and the sample units. These adjustments may include the creation of separate high-growth strata and the specification of rules for handling changes in administrative divisions that affect sampling units or strata. Although this revision strategy has been used in the Australian master sample, it seems to be rarely used in developing countries. One reason is probably that this strategy is complex from a sampling point of view, requiring greater care and skill in design and execution.

6. Documentation

87. Much of the documentation work is already done if the master sample has been selected from a well-documented master sampling frame. Documentation, however, is sometimes a weak aspect of master samples in developing countries. The information may be scattered and sometimes scarce, making it difficult to follow the selection of the sample and to calculate sampling probabilities. The selection procedures and the selection probabilities for all of the master sample units at every stage must be fully documented. There should also be records showing which master sample units have been used in samples for particular surveys. A standard identification number system must be used for the sampling units.

88. The documentation of the master sample should also include measures of master sample performance in terms of sampling errors and design effects for important estimates. These performance measures are useful for the planning of sample sizes and sample allocation in new surveys based on the master sample. Procedures for calculation of correct variances and design effects are now available in many statistical analysis software packages (see chap. XXI for details).

89. The documentation should also include auxiliary materials for the master sample. If secondary sampling frames (SSF) have been prepared for the master sample USUs, then these frames should be part of the documentation. The SSFs will consist of area units such as blocks or segments or of list units such as dwelling units within the master sample USUs.

7. Using a master sample for surveys of establishments

90. The main purpose of a master sample is to provide samples for the household surveys in the continuous survey programme (and any ad hoc survey that fits into the master sample design). The sample will thus primarily be designed to serve a basic set of household surveys. It will generally not be efficient for sampling of other types of units. In some situations, however, it may be possible to use the master sample for surveys concerned with the study of characteristics of economic units, such as household enterprises, own-account businesses and small-scale agricultural holdings.

91. In most developing countries, a large proportion of the economic establishments in the service, trade and agricultural sectors are closely associated with private households. Those establishments are typically many in number and small in size and they are widely spread throughout the population. There may often be a one-to-one correspondence between such establishments and households, and households rather than the establishments themselves may serve as the ultimate sampling units. A master sample of households can be used for surveys of these types of establishments. This will often require departures from self-weighting designs. Verma (2001) discusses ways of improving the efficiency of sample design for surveys of economic units.

92. There are, however, usually a number of large establishments that are not associated with households. These establishments are typically rather few but they account for a large proportion of many estimates of totals (output, number of employees, etc.). They are also, in many cases, unevenly distributed with respect to the general population. As the master sample of areas will not sample these large units in an efficient way, a separate sampling frame is needed for them. In many cases, such a frame could be constructed from records of government agencies (for example, taxation or licensing agencies). From this list, all of the very large units and a sample of the remaining units should be selected for the survey, along with a sample of establishments from the master sample PSUs.

93. A special case of an establishment survey arises when a household survey is linked to a "community survey". For example, in a health survey, the survey of individuals/households may be supplemented by a survey of health-care facilities covering extended areas around each of the original sample areas (for example, enumeration areas). Data from the supplementary survey may have two purposes: (a) it can be linked to the household data and used for analyses of the quality and accessibility of local facilities; and (b) it can be used to produce national estimates of the number and types of health facilities. For the first purpose, the households/individuals remain the unit of analysis: no new sampling issues are involved. The second purpose can produce more complications. If the larger extended area around the original sample area is taken as a larger unit (district, commune, census supervision area, etc.) consisting of a number of areas along with the sampled area, then the situation is simple. The resulting sample would be the equivalent of a sample of larger areas with the probability of selection of the larger area equal to the sum of selection probabilities for the smaller areas contained within the larger area. If, however, the larger area is constructed by the rule "within x kilometres of the original sample area", the determination of selection probabilities is more complex.

E. Concluding remarks

94. The design and execution of household surveys is an important task for all national statistical offices. Many NSOs in developing countries carry out several surveys every year. The need for the planning and coordination of the survey operations has stimulated efforts to integrate the surveys in household survey programmes. The idea of an integrated household survey programme is now being realized in many national statistical offices.

95. An important part of the work with a survey programme is the design of samples for the different surveys. This chapter has addressed the key issues concerning the design and

development of master sampling frames and master samples. The advantages of a well-kept master sampling frame have been described and it has been argued that every NSO executing a household survey programme should have a well-kept master sampling frame that could cater for the needs of the household surveys in the survey programme and also for the needs of ad hoc surveys that may crop up during the survey programme period. Furthermore, many NSOs can go a step further and design and use a master sample for all or most of the surveys in the survey programme and possibly for unanticipated ad hoc surveys.

96. The chapter has given an overview of the important steps to be taken when developing master sampling frames and master samples and has provided illustrations of master sampling frames and master samples from some developing countries. Its format does not allow for a detailed treatment of all the important issues related to the development of master sampling frames and master samples. Readers who would like a more thorough description should consult the relevant United Nations manual (see United Nations, 1986).

References

Ajayi, O.O. (2000). Survey methodology for the sample census of agriculture in Nigeria with some comparisons of experiences in other countries. Paper presented at the International Seminar on China Agricultural Census Results held in Beijing, 19-22 September 2000.

Glewwe, P., and I.Yansaneh (2001). *Recommendations for Multi-Purpose Household Surveys from 2002 to 2010.* Report of Mission to the General Statistics Office, Viet Nam.

Kish, L. (1965). *Survey Sampling.* New York: John Wiley and Sons.

_____1988). Multi-purpose sample design. *Survey Methodology*, vol. 14, pp. 19-32.

Pettersson, H. (1994). *Master Sample Design: Report from a Mission to the National Central Statistics Office, Namibia, May 1994*. International Consulting Office, Statistics Sweden.

_____ (2001a) *Sample Design for Household and Business Surveys: Report from a Mission to the Bureau of Statistics, Lesotho, 21 May – 2 June 2001*. International Consulting Office, Statistics Sweden.

_____ (2001b). *Recommendations Regarding the Design of a Master Sample for the Household Surveys of GSO: Report of Mission to the General Statistics Office, Viet Nam*. International Consulting Office, Statistics Sweden.

Rosen, B. (1997). *Creation of the 1997 Lao Master Sample: Report from a Mission to the National Statistics Centre, Lao PDR*. International Consulting Office, Statistics Sweden.

Torene, R., and L.G. Torene (1987). The practical side of using master samples: the Bangladesh experience. *Bulletin of the International Statistical Institute: Proceedings of the 46th Session, Tokyo, 1987*, vol. LII-2, pp. 493-511.

United Nations (1986). *National Household Survey Capability Programme: Sampling Frames and Sample Designs for Integrated Household Survey Programmes (Preliminary Version*). DP/UN/INT-84-014/5E, New York.

Verma, V. (2001). Sample design for national surveys: surveying small-scale economic units. *Statistics in Transition*, vol. 5, No. 3 (December 2001), pp. 367-382.