INTEGRATED SURVEY FRAMEWORK

GUIDELINES
Publication prepared in the framework of the Global Strategy to improve Agricultural and Rural Statistics
Cover photos:
©PATH/Gabe Bienczycki
©CGIAR/P. Casier
Guidelines for the Integrated Survey Framework
# Table of Contents

ACRONYMS ......................................................................................... 8
ACKNOWLEDGEMENTS ..................................................................... 9
PREFACE .......................................................................................... 10

1 Introduction .................................................................................... 12
   1.1 Introduction ................................................................................ 12
   1.2 Overview of the Proposed Method .............................................. 17
   1.3 Reasons for the Method Proposed .............................................. 20
   1.4 The Context and Other Data Integration Initiatives within FAO ...... 22
   1.5 Guidance for the Reader .............................................................. 23

2 Basic methodology ........................................................................ 25
   2.1 Introduction ................................................................................ 25
   2.2 Statistical units of interest ............................................................ 26
   2.3 The relationships between units ................................................ 28
   2.4 Observational methodology ....................................................... 34
   2.5 Estimation .................................................................................. 38
   2.6 A note on the estimation of sampling errors ............................... 45
   2.7 Special case: subsampling the households of large farms ............ 46

3 Modules and operational rules for the application of indirect sampling in agricultural surveys ................................. 49
   3.1 Overview ................................................................................... 49
   3.2 The conceptual definition of links .............................................. 50
      3.2.1 Issues .................................................................................. 50
      3.2.2 Links between households and farms ................................. 52
      3.2.3 Level of link definition ......................................................... 55
Part I – Modules and operational rules for observing farms starting from households ................................................................. 56

1 Collecting data on the links .......................................................................................................................................................................................... 56
   1.1 Identification and codification of farms .................................................................................................................................................. 57
   1.2 Collecting data on the links of observed farms ................................................................................................................................. 58
   1.3 Subsampling during the observation of farms ................................................................................................................................. 59

2 A practical example ......................................................................................................................................................................................... 60
   2.1 Description of the study ................................................................................................................................................................. 60
   2.2 Choice of the household frame .................................................................................................................................................. 60
   2.3 Selection of the household sample ............................................................................................................................................... 61
   2.4 Choice of links ........................................................................................................................................................................... 61
   2.5 Collecting data on the links .......................................................................................................................................................... 62
      2.5.1 Questionnaire on household links ........................................................................................................................................ 62
      2.5.2 Site links .................................................................................................................................................................................. 63

3 Simulation of the estimation of farms’ weights ........................................................................................................................................... 64
   3.1 Recalling the method of weight estimation using the GWSM ................................................................................................. 64
   3.2 A numerical example ................................................................................................................................................................. 65
   3.3 Excel simulation ............................................................................................................................................................................ 66

Part II – Observing households starting from a survey on farms ........... 69

1 Definition of links ........................................................................................................................................................................................................... 69

2 Data collection on the links ........................................................................................................................................................................... 70
   2.1 Number of links of each farm .................................................................................................................................................. 70
   2.2 Subsampling of the households observed ........................................................................................................................................ 71
   2.3 The total number of links of observed households .................................................................................................................... 71

3 A practical example ......................................................................................................................................................................................... 72
   3.1 Choice of link .................................................................................................................................................................................. 72
   3.2 Data collection on links ................................................................................................................................................................. 72
   3.3. Weight calculation ........................................................................................................................................................................ 74
   3.4 Simulation with Excel data ......................................................................................................................................................... 76
4 Changes in statistical units

4.1 Introduction

4.2 Analysis of changes (households and farms)
   4.2.1 Disappearance
   4.2.2 Fusion
   4.2.3 Division
   4.2.4 Population change

4.3 The main solutions for correcting sample changes
   4.3.1 Additional samples
   4.3.2 Oversampling
   4.3.3 Post-stratification
   4.3.4 Tracking

4.4 Changes in households and weight calculation
   4.4.1 Classic methods of weight-sharing
      4.4.1.1 Equal household weighting scheme
      4.4.1.2 Equal person weighting scheme
      4.4.1.3 Application
         4.4.1.3.1 Calculation of weight with an equal household weighting scheme
         4.4.1.3.2 Calculation of weight with the equal person weighting scheme
   4.4.2 The General Weight Share Method
      4.4.2.1 Definitions of links
      4.4.2.2 Collecting data on the links
      4.4.2.3 Application

4.5 Changes in farms and weight calculation
   4.5.1 Equal farm weighting scheme
   4.5.2 Equal plot weighting scheme
   4.5.3 The General Weight Sample Method
      4.5.3.1 Definitions of links
      4.5.3.2 Collecting data on the links
Acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGRIS</td>
<td>Agricultural and Rural Integrated Survey</td>
</tr>
<tr>
<td>AH</td>
<td>Agricultural Holding</td>
</tr>
<tr>
<td>CV</td>
<td>Coefficient of Variation</td>
</tr>
<tr>
<td>FAO</td>
<td>Food and Agricultural Organization</td>
</tr>
<tr>
<td>EA</td>
<td>Enumeration Area</td>
</tr>
<tr>
<td>GS</td>
<td>Global Strategy to Improve Agricultural and Rural Statistics</td>
</tr>
<tr>
<td>GWSM</td>
<td>Generalized Weight Share Method</td>
</tr>
<tr>
<td>HH</td>
<td>Household</td>
</tr>
<tr>
<td>ISF</td>
<td>Integrated Survey Framework</td>
</tr>
<tr>
<td>MM</td>
<td>Multidomain and Multivariate method</td>
</tr>
<tr>
<td>MSCD</td>
<td>Minimum Set of Core Data</td>
</tr>
<tr>
<td>MSF</td>
<td>Master Sampling Frame</td>
</tr>
<tr>
<td>NEPS</td>
<td>National Educational Panel Study (Germany)</td>
</tr>
<tr>
<td>NGO</td>
<td>National Governmental Organization</td>
</tr>
<tr>
<td>PSU</td>
<td>Primary Sampling Unit</td>
</tr>
<tr>
<td>PPS</td>
<td>Probability Proportional Size</td>
</tr>
<tr>
<td>RGA</td>
<td>General Census of Agriculture (Burkina Faso)</td>
</tr>
<tr>
<td>RGPH</td>
<td>General Census of Population and Housing (Burkina Faso)</td>
</tr>
<tr>
<td>SAS</td>
<td>Statistical Analysis Software</td>
</tr>
<tr>
<td>SLID</td>
<td>Survey of Labour and Income Dynamics (Canada)</td>
</tr>
<tr>
<td>SSRS</td>
<td>Stratified Simple Random Sampling</td>
</tr>
<tr>
<td>WCA</td>
<td>World Census of Agriculture (FAO)</td>
</tr>
</tbody>
</table>
Acknowledgements

These Guidelines for the Integrated Survey Framework have been developed as a specific topic of the “Technical Report on the Integrated Survey Framework”, within the Research activity carried out by the Global Strategy to improve Agricultural and Rural Statistics. This work was prepared under the guidance of Naman Keita, Senior World Bank/FAO Consultant and Interim Coordinator of Research of the Global Office, and coordinated by Piero Demetrio Falorsi, Senior Consultant, who has supervised the research team in different areas:

Angela Piersante – Introduction, Purpose and Scope of the Guidelines
Piero Demetrio Falorsi – Methodology of Indirect Sampling
Bako Dramane – Analysis and Presentation of the Model’s Application in Different Scenarios, Using Examples and Practice from Burkina Faso
Paolo Righi – Application of Optimal Sampling on Countries’ Data

Pierre Lavallée, Senior Statistician at Statistics Canada, and Michael N. Isimwaa, Deputy Director of Policy and Planning Department (Ministry of Agriculture and Livestock of Zambia) kindly provided their generous and constructive collaboration as peer reviewers. In particular, we wish to thank Mr Lavallée for his precious contributions to the parts concerning the practical application of his innovative methodology on indirect sampling, and Mr Isimwaa for kindly making available to us his expertise and invaluable experience in the field of agricultural and rural statistics.

The document was edited by Sarah Pasetto, FAO/Global Strategy.
Preface

The growing demand by policymakers and decision makers for statistics based on information that is interlinked in economic, social, and environmental aspects requires the large-scale expansion of national efforts to implement statistical surveys, in terms of organization and budget. Therefore, the collection of data by integrating information from different sources is becoming a crucial requirement for the production of statistics.

**Agricultural censuses** are performed only periodically and at long intervals of time. This makes them unsuitable for providing the timely and regular information that is required for the effective monitoring and evaluation of development programmes' impact on agriculture and the rural economy.

On the other hand, **sample surveys** are usually carried out with a greater frequency, but the data are collected by sector, using different sampling frames and without any possibility of measuring the cross-sector impact of a given action; in addition, the sample size is usually too small for the data to be amenable to disaggregation into lower geographical areas or rural and farm sectors.

Furthermore, the organizational approach adopted in certain countries with regard to agricultural sample surveys does not allow for the optimization of methods, costs and time. Very often, **ad hoc** surveys that collect the same final information have different objectives and are performed by different national agencies, using separate budgets and methods; thus, different findings often result.

However, agricultural surveys do not only collect economic data. They can also be used to expand the information that is available on the agricultural sector, by integrating demographic data and thus enabling a more complex social analysis.

The integration of holding and household data is a new approach that can be adopted when conducting sample surveys. It enables provision of timely and cost-effective methods – with a measurable level of reliability – for monitoring and evaluating the impact of development programs on various economic sectors and socio-environmental settings. Indeed, the integrated observation of households and holdings in a single survey is very complex to achieve with traditional sampling techniques, except in the case of simple (one-to-one) relationships between households and holdings. These Guidelines explain how to deal with this complex topic with reference to recent developments in the field of indirect sampling techniques.
In particular, these Guidelines illustrate how to apply the “observation” method based on indirect sampling, to guide countries in building and strengthening national capacities for collecting data in such a way that responds to standard data quality requirements. The Guidelines are intended especially for use by the staff of national statistical offices and ministries or agencies that are involved in data collection on agriculture, and seek to provide basic information with a step-by-step approach, beginning with practical examples drawn from the experiences of certain countries.

These Guidelines illustrate a statistical method that produces correct estimates with reduced time and costs; in particular, they demonstrate the application of indirect sampling to integrated household surveys for a wide range of topics on the basis of the correspondence between households and holdings, so that these may be observed in various scenarios within a single survey.
Introduction

1.1 Introduction

The growing demand by policy-makers and decision makers for statistics based on information that is interlinked in economic, social, and environmental aspects requires a large-scale expansion of national efforts to implement statistical surveys, in terms of organization and budget. Therefore, the collection of data by integrating information from different sources is becoming a crucial requirement for the production of statistics.

In developing countries, agriculture plays a key role in overall economic growth, increasing incomes, reducing poverty and fighting hunger. Within the sector, agricultural censuses are the largest source of statistical information, due to their wider geographical coverage that includes most national farms and holdings. Reliable and comprehensive statistics are a basis for the formulation, monitoring and evaluation of national agricultural development plans. Therefore, the collection of reliable and relevant data on the food and agricultural sector is an important process for providing baseline information and statistics for the proper planning and monitoring of agricultural programmes aimed at achieving food and nutrition security, employment creation, income generation and poverty reduction.

Agricultural censuses are not the only source of food and agricultural data. Indeed, other sources – such as population censuses, administrative reports, and household sample surveys – may not specifically cover the entire agricultural sector, but may still provide relevant information. Thus, the information sought should be the result of a
statistical system or of a combination of different sources, which are linked to each other and share a common conceptual and methodological basis, or at least mechanisms to foster complementarities. The type or combination of methodologies to be used in the generation of food and agricultural statistics should be determined by the purpose or objectives for which the information is required, as well as the budgetary resources available and time constraints.

Data collection activities vary across countries in terms of local items, periodicity and methods; however, the data on agriculture can be organized as follows:

1) Current data. These relate to agricultural activities that are almost continuous and are repeated every year. Examples are crop area, yield and production of crops and livestock, production inputs, utilization of output, and prices. Usually, these data are collected through sample surveys on a continuous or seasonal basis, possibly several times in an agricultural year.

2) Structural data. These reflect the structure of the country’s agricultural economy, reporting elements such as the number of holdings, machinery, manpower, land cover and use. Since changes in this context generally do not occur very rapidly, this information need not necessarily be compiled on a frequent basis; compilation every five or ten years is sufficient. These data are usually collected through agricultural censuses.

Although comprehensive, agricultural censuses are performed only periodically and at long intervals of time. This makes them unsuitable for providing the timely and regular information that is required for the effective monitoring and evaluation of development programmes’ impact on agriculture and the rural economy.

On the other hand, sample surveys are usually carried out with a greater frequency and should be capable of providing more detailed statistical information, because the sizes of the sample surveys can be adjusted to take into consideration acceptable levels of minimum standard errors of the resulting estimates, in relation to the size of the population of interest. The sample size should also take into account the need by other users and stakeholders for reliable estimates at lower geographical units, such as province, district, and constituency and ward levels, in addition to national level estimates.

However, the data are often collected by sector, using different sampling frames and without any possibility of measuring the cross-sector impact of a given action; in addition, the sample size is sometimes too small for the data to be amenable to
disaggregation into lower level geographical units and farm sectors, and the model accuracy of small area estimates may be weak.

In addition, the organizational approach adopted in certain countries for agricultural sample surveys does not allow for optimization of methods, costs and time. Very often, *ad hoc* surveys that collect the same final information have different objectives and are performed by different national agencies, using separate budgets and methods; thus, different findings often result.

The examples below were taken from a national report on the annual agricultural surveys programme. They show that although the holding was the common observation unit of the three objectives, various agricultural sample surveys were carried out during the year in question:

- **Objective of Survey A**: “estimating number of major livestock by age and sex in each season, and annual production of milk, eggs and meat with a percent Coefficient of Variation (CV) around 10% and collect reliable data on feeding and management practices of livestock”.

- **Objective of Survey B**: “estimating production of paddy and maize in the country with an accuracy corresponding to CV around 3% and obtain reliable estimate of consumption of chemical fertilizers”.

- **Objective of Survey C**: “obtaining reliable estimates of area under high yielding and improved varieties of cereal crops and extent of use of modern cultivation practices in province X”.

A single integrated survey could incorporate all these objectives into a common data collection process, dealing with the same observation unit on basic agricultural items.

However, agricultural surveys do not only collect economic data. They can also be used to expand the information that is available on the agricultural sector, for example by integrating:

- Demographic data. This would enable a more complex social analysis. For example, assuming that the contribution of women’s work to agriculture is to be evaluated, a sample survey should jointly analyse the household sector – to collect demographic data – and the holding sector, to collect data on the land tenure from which the female household members generate annual income. The relationship between the two sectors is defined by the various roles that women perform within the holding. In this case too, a single integrated survey could optimize data collection to achieve a more comprehensive data analysis.
Agricultural management data on e.g. livestock practices, so that they may not only supplement food supply (milk, eggs, and meat) but also enable the provision of off-season and part-time employment for households and raw material for industries (wool, hides, skins, hair, bristles, etc.), thereby making a significant contribution to national income. In planning a production development that also responds to environment mitigation purposes, the data collected on the yields of milk, eggs, meat, feeding, labour and management practices are all necessary; thus, the adoption of an integrated survey may be the best solution.

Moreover, the use of administrative data for generating rural and development indicators is not yet common in all countries, although its importance is growing especially in the production of income data for farm operators.

Consequently, the results of data collection based on current practices often conflict, and are not integrated into a common framework that makes data easily accessible and comparable; this means that implementation efforts in terms of costs, time and the efficient use of resources (budget, people, training, etc.) are greater.

The collection of data by integrating it into a common framework could facilitate countries’ efforts to produce coherent and comparable data by using standards, analysing data across sectors and improving the sustainability of countries’ statistical capacities.

Indeed, field experience in developing countries has shown that while structural agricultural data can be obtained from government-managed farms and sometimes from large private farms, the traditional sector – or subsistence agriculture – constitutes a very high proportion of the total agricultural sector in which households are engaged. The traditional or smallholder sector households are simultaneously units of production, revenue generation and consumption. Thus, in this sector, the phenomena under consideration can be properly understood only if holdings and households are observed jointly. Moreover, a “one-to-one” correspondence can often be assumed to exist between households and traditional holdings, such that collecting data on traditional agriculture holdings is equivalent to collecting data on households.
However, this is not always the case. In the traditional sector too, more complex relationships between these two entities can be observed. An example is that of cooperatives, in which many households contribute to the same farm; another common case is the situation in which the head of the household operates one farm, while other members are employed in other large farms.

The integration of holding and household data is a new approach that can be adopted when conducting sample surveys, which enables the provision of timely and cost-effective data – with a measurable level of reliability – to monitor and evaluate the impact of development programs on various economic sectors and within different socio-environment settings. Indeed, the integrated observation in a single survey of households and holdings is very complex to achieve with traditional sampling techniques, except in the case of simple (one-to-one) relationships between households and holdings\(^1\). These Guidelines explain how to deal with this complex topic, with reference to some recent developments in indirect sampling techniques.

In particular, with the ultimate aim of supporting countries in building and strengthening national capacities for data collection that respond to standard data quality requirements, these Guidelines seek to deliver indications on applying the “observation” method based on indirect sampling. They are especially intended for use by the staff of national statistical offices and Ministries of Agriculture or Agencies involved in data collection on agriculture. These Guidelines seek to provide basic information with a step-by-step approach, beginning with practical cases drawn from the experiences of certain countries.

\(^1\) Indeed, some national practices show that countries or communities experience issues related to holdings that are commonly owned by different households and to households owning multiple holdings, some of which are located at a great distance from one another and in unselected enumeration areas. These aspects must be taken into consideration when constructing the integrated sampling frame (households and holdings) and the resulting sampling of units to be selected as survey targets.
These Guidelines illustrate a statistical method that produces correct estimations with reduced time and costs; in particular, they display the application of indirect sampling to integrated household surveys for a wide range of topics on the basis of the correspondence between households and holding, so that they may be observed in various scenarios, within a single survey.

1.2 Overview of the Proposed Method

Traditional survey theories include non-probability sampling and probability sampling. The latter is carried out by randomly selecting sample units to establish the probability of drawing certain samples from lists called sample frames. These sampling frames represent the set of observation units that produce the information on the phenomenon under analysis.

Very often, the survey refers to the units or individuals that are related to each other by economic and social activities, and are regulated by formal rules, contingent dependencies or relationships created for the pursuit of common purposes. These relationships represent groups of a given social setting that are interlinked in different aspects and that also include individuals belonging to other groups.

The individuals belonging to a population that is the object of a given survey become carriers of information on another statistical population, through the type of relationship between the entities. Observation of this link and a representativeness analysis allow for a methodological approach based on indirect sampling to be proposed.

Indirect sampling provides a framework for the estimation of the parameters of two target populations that are related to each other through the observation units: a framework that is based on a joint analysis of this relationship. One advantage is the choice of a sampling frame that can be indirectly related to the population for which the frame is incomplete, unavailable or not up-to-date; in addition, data can be obtained from observation units that are related to each other.

This method’s sampling strategy lies in the sampling design, the calculation of the estimator and the observational process.
Figure 1.2 below is an example of a simplified scenario of all the possible links between two populations that are related to each other because the units under observation – the household and the holding – correspond.

1) The relationships, which are unknown before the observation is analysed, derive from the economic agricultural activities performed by members of the household, and can be one-to-one (Cases a and d), many-to-many (Case b) and one-to-many (Case c).

2) The rectangles represent the holdings (clusters), which include the head of the holding and its workers; the latter could belong to different households. Each holding or cluster includes all the links or workers within the household population.

Assuming that the population of households within a frame is available and that the holdings population is the related population, the main phases of indirect sampling are the following:

1) selection of the probability sample (e.g. using a traditional multistage proportional-to-size – PPS – stratified design) from the population for which a frame is available, that could be the household population (e.g. Households a and c), and calculation of the inclusion probability for each unit;
2) identification of the sample of the population without a frame available, which is the holding population; analysis and observation of the existing relationships (links) between the two populations, and submission of any specific questions that may arise.

3) assignment of a weight to each unit of the holding population, applying the Generalized Weight Share Method (GWSM).

The weight estimation of each surveyed unit of the holding population using the GWSM is obtained by calculating the average of the sampling weights of the household population from which the sample is selected. On an operational level, this entails:

   a. identification of the links between each household and the holding, by interviewing the household population sampled (for each household sampled, the reference of the holding in which household members perform agricultural activities must be provided) – for example, if the sample includes only Households a and c, the corresponding Holdings are I, III, and IV;

   b. identification of all links between the holding and the household population, by interviewing each holding (providing the number of all workers who participate in the holding’s activities, including household members that were not selected to build the sample);

   c. calculation of the weight for each surveyed unit of the holding population, by computing the average of the sampling weights of the household population from which the sample is selected (dividing the sampling weights of the household members who work for a given holding by the total number of the links or workers belonging to the holding).

The number of links should be provided during the survey interview of the holding population in a specific section, to identify the population without an available frame. A questionnaire based on a face-to-face method with a semi-fixed structure should be used. The enumeration should pose additional questions (probing) that were not included in the original format, to provide and obtain further information that can facilitate comprehension of the relationships, obtain the interviewer’s collaboration, build a certain degree of confidence, verify that the response is appropriate, assist the respondent if the question is difficult to understand, etc.
4) estimation of the holding population’s parameters, such as totals and means, with the weights obtained from use of the GWSM.

This sampling method is flexible enough to be applied on the basis of various frames; however, strong technical skills are required to address the questions on the social and economic factors that can define the relationships between the two populations.

1.3 Reasons for the Method Proposed

These Guidelines illustrate how to exploit a survey on a given target population (e.g. the population of households) to produce information that is relevant to agricultural and rural statistics. This enables the collection of information to be expanded to include the collection that can now be performed on populations of statistical units other than those of the target population (e.g. holdings).

This aspect has relevant practical implications, because many developing countries have the statistical capacity to regularly carry out at most one survey; the same sample can thus be used to collect data on a more complete statistical population, for the purposes of agricultural and rural statistics.

The added value of this method is its capacity to broaden the range of information in ways that were not envisaged in the initial objectives; it can be used as an option for the development of a survey that has already been structured. This method could be very useful to broaden the scope of other projects undertaken by international organizations (e.g. the Core Agriculture and Rural Data Survey, the Living Standards Measurement Survey, etc.) that deal with multipurpose surveys on households and holdings.

Indeed, the indirect sampling approach proposed can be considered a statistical analysis tool that is applicable to any sample, to provide additional information that is not obviously necessary at an initial analysis, but that emerges as such upon closer consideration due to the cross-relations between the units. The approach uses statistical procedures that are based on the representativeness of the original sample to provide consistent and accurate statistics, and has the following advantages.

1. Harmonizing sample frames in building a master sampling frame for all modules.

For example, in conducting an agricultural census linked with the population census, one country followed the practice of using the household frame for basic modules (rain-fed crops, livestock and arboriculture) only, while the other
specific modules, on fisheries (observation unit: fishermen) and irrigated crops (observation unit: irrigated parcel holders) were conducted with separate frames. The module on modern agricultural holdings was conducted through an exhaustive survey on modern agricultural holdings.

2. The same sampling frame can be used to develop surveys that are conducted separately and are based on different observation units (e.g. a specific thematic survey on farms concerning the use of fertilizers, or a survey on households regarding demographic topics);

3. Data surveys can be harmonized to provide coherent, consistent and comparable data, as common concepts, definitions and classifications are adopted in relation to the observation units (i.e., holdings and households), the basic items and the data set to be collected. For example, with regard to the data on meat production for human consumption, it is necessary to collect data on the total number of slaughtered animals, but several countries collect data on controlled slaughters on the basis of reports from slaughterhouses. However, in most developing countries, non-controlled slaughters of certain animals are considerably greater than controlled slaughters, so that the meat supply for human consumption cannot be estimated. The adoption of common standards raises the need to develop in-depth survey modules on the subject, to provide the statistics that are required but are not exhaustively provided with the current means.

4. The use of sampling frames that have become obsolete, to produce accurate estimations through statistical procedures.

5. The use of different frames, master frames or other sample frames due to the framework’s flexibility.

6. A more efficient use of resources, enabling, for example, the reduction of the cost and time required for implementation and for technical personnel training.
1.4 The Context and Other Data Integration Initiatives within FAO

The methodology proposed herein is a particular approach to solving the problems of an integrated observation of households and farms, and is a result of a research project undertaken by the Global Strategy to improve Agricultural and Rural Statistics. The main results are explained in the Technical Report on the Integrated Survey Framework, published by the Global Strategy and available online at http://www.gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf.

The Technical Report sets out in detail the various techniques for ensuring an integrated observation. The volume is divided into five main parts, which refer specifically to:

- **Part 1.** Purpose and FAO experiences
- **Part 2.** International standard requirements on data quality frameworks and the practices followed by certain countries in relation to the quality of integrated survey frameworks (the quality issues of an integrated observation)
- **Part 3.** Record linkage (how to exploit a record linkage procedure with data referring to different statistical units)
- **Part 4.** Sampling (how to construct a sample for the joint observation of farms and households, considering a multiple frame context)
- **Part 5.** Estimation (how to achieve coherent estimates when different surveys produce conflicting estimates relating to the same phenomena).

In particular, all technical results used in these Guidelines were derived in Part 4 of the Report.

The Technical Report forms part of the methodological material proposed by the Global Strategy’s Research program to develop an Integrated Survey Framework.

The implementation of the Integrated Survey Framework enables agricultural statistics to be integrated into national statistical systems, together with:

- the development of a Master Sample Frame for agriculture (consistent with the initiative of linking population and agricultural censuses), which should be the source of all samples for surveys of agricultural holdings, farm households and rural non-farm households; and
- the design of an Integrated Management Data System that can describe the integration during the phase of data dissemination.

The Integrated Survey Framework is implemented by another initiative that is crucial to integration, proposed by FAO under the aegis of the Global Strategy: the Agricultural and Rural Integrated Survey (AGRIS). This initiative proposes an integrated survey option to plan the collection of data on agriculture in accordance with a standard structure, so that all countries can provide a Minimum Set of Core Data that meets current and emerging demands relating to national development policies.

The AGRIS methodology is based on a standardized multipurpose and modular survey, to be submitted to agricultural holding sectors on a regular basis. The modules are organized as follows:

a) one Core module, to collect yearly current data on agricultural production, economic, and socio-demographic statistics;

b) standard modules on specific topics (Economic=M1, Demographic=M2, etc.) to collect other data that are more stable and require a variable periodicity, e.g. every two, three or five years.

Indirect sampling is highly flexible and thus suitable for this setting’s specific purposes, because it can be adapted to various options by using a Master Sampling Frame (if one has been developed) or a Sample Frame (population/agricultural census enumeration areas, household registers from population censuses, registers of farms from agricultural censuses, registers of farms based on administrative records, area sample frames, or multiple frames). Therefore, the relationship with the agricultural census, and in particular with the integrated modular approach of the World Census of Agriculture, is fundamental to determining the frame. The frames provide a basis for the selection of a probability sample of holdings or households that can link units in various relationships, and for a conceptual framework that is coherent with the concepts and definitions adopted.

1.5 Guidance for the Reader

These Guidelines deal with the main issues relating to the construction of an integrated strategy for the joint observation of households and holdings, for the ultimate purpose of producing unbiased statistics on households and farms. They will focus on how to perform these steps, taking into account practical examples from the developing countries involved in developing this initiative. The specific topics examined in the following chapters are illustrated briefly below.
Chapter 2 provides a detailed explanation of the overall approach. It discusses:

- the definitions of the agricultural setting’s observational units, the population of households and the population of farms/holdings;
- the possible relationships between the two populations;
- the integrated observation method;
- the construction of the target population’s weights and the calculation of sampling errors.

Chapter 3 analyses all the phases of the model’s application, using data and scenarios from the experience of Burkina Faso, beginning with the household frame to ascertain the relation between the populations of households and of farms/holdings. It also analyses the inverse chain: how to observe households, starting from farms. For both situations, it includes an analysis and definition of the links between the two populations, and proposes a model questionnaire to observe and build the weights in the different scenarios. Examples and simulations from the Burkina Faso data are proposed.

Chapter 4 analyses the issue of the change of statistical units in the populations of both households and farms, and provides a description of the changes intervening between the initial observation unit and the new one, depending on the following phenomena: merger, division and creation of a new unit. This chapter also suggests an analysis of the links related to the changes, with insight on the various advantages and disadvantages, including questions to investigate the changed scenarios.

Chapter 5 focuses on how to define an optimal sample for a survey that seeks to produce an integrated observation of both household and farm populations. The optimality criterion is the minimum cost solution that enables estimates to be produced for both populations with predefined levels of accuracy.
Basic methodology

2.1 Introduction

This Chapter establishes a conceptual framework for integrated observation in the agricultural sector. It includes common definitions adopted for the relevant statistical units and sets out the basic concepts of the methodology proposed, which is based on a joint use of indirect sampling (Lavallée 2007) and multiple frame sampling (Singh and Mecatti 2011) and is illustrated in detail the Technical Report on Integrated Survey Framework (Part 4).

The methodology is based on a framework that meets the following fundamental requirements:

- a social relationship (economic activities and/or demographic status) exists between populations that appear to belong to different settings, such as households and holdings;
- the relationships (or links) between the members of the two populations are analysed;
- a frame for one of the two populations is available.

The implementation process features the following components:

- an observational phase, focusing on the links between a population for which a frame is available and another population with an unknown frame;
- a probabilistic sampling strategy, to be applied to the population with the available frame and that requires that:

  a) all units in the population be represented in the frame used to select the sample
  
  b) the selections be made by applying a proper randomized procedure, which assigns definite selection probabilities to all eligible units
  
  c) the sample, as selected, be successfully enumerated in the survey

- the parameters’ estimation of the related population, through statistical procedures and error estimation.

The topics will be examined in detail to analyse the various scenarios existing in household and holding populations, the links between the individuals belonging to household and holding clusters, and the estimates obtained using the weights derived from one of the two populations.

Through practical examples and on the basis of an assumed frame (the sampled population with its inclusion probability), the observational method can be divided into two phases: first, the first population (e.g. households) is observed; then, the units of the other population (e.g. farms) are surveyed, using an indirect sampling mechanism that exploits the links with the units of the first population sampled.

### 2.2 Statistical units of interest

In this context, the observation unit is one of the main issues to be addressed. For other types of survey within the national survey program, the observation unit is the household; for agricultural surveys, the observation unit is an agricultural holding. Thus, the statistical units that play a key role within integrated surveys are the household and the agricultural holding:

- **Household**: this is based “on the arrangements made by persons, individually or in groups, for providing themselves with food or other essentials for living. A household may be either 
  (a) a one-person household, that is to say, a person who makes provision for his or her own food or other essentials for living without combining with any other person to form part of a multi-person household, or
  (b) a multi-person household, that is to say, a group of two or more persons living together who make common provision for food or other essentials for living. The persons in the group may pool their incomes and may, to a greater or lesser extent, have a common budget; they may be
related or unrelated persons or constitute a combination of persons both related and unrelated” (United Nations 2008, p. 100).

- Agricultural Holding: “An agricultural holding is an economic unit of agricultural production under single management comprising all livestock kept and all land used wholly or partly for agricultural production purposes, without regard to title, legal form, or size. Single management may be exercised by an individual or household, jointly by two or more individuals or households, by a clan or tribe, or by a juridical person such as a corporation, cooperative or government agency. The holding’s land may consist of one or more parcels, located in one or more separate areas or in one or more territorial or administrative divisions, providing the parcels share the same production means, such as labor, farm buildings, machinery or draught animals.” (FAO 2010).

The head of the household and the agricultural holder are the parties that can provide information for, respectively, the entire unit of the household and the farm. The agricultural holder “is defined as the person who makes the major decisions regarding resource use and exercises management control over the agricultural holding operation” (FAO 2010).

For all practical purposes, the concept of Agricultural Holding (AH) coincides with that of Farm. In these Guidelines, the two terms are used interchangeably. However, we note that the term “Agricultural Holding” tends to be adopted in agricultural censuses, while that of “Farm” is more frequently used in surveys on agriculture.

Two categories of AH/Farm are involved in agricultural production. According to the FAO World Program Census (WCA), these can be defined as follows:

- **AHs in the household sector** – in this case, the enterprise is the agricultural production management unit in the household, and the agricultural holding is the unit of agricultural production within the management unit.
- **AHs in the non-household sector** – these are mainly large farms or cooperatives.

The two categories define two separate sub-populations, which constitute a partition of the total population of AHs.
2.3 The relationships between units

Both households and farms are complex statistical units that may be viewed as clusters of elementary units. Indeed,

- Households can also be considered clusters of individuals. Each individual belongs to only one household.

- Farms can also be viewed as:
  - a set of individuals (workers or farm holders). There is an important difference between households and farms: while each individual belongs to only one household, an individual may be linked to more than one farm. For example, the individual may be the agricultural holder of a farm in the household sector and at the same time a worker of a farm in the non-household sector. However, the most common situation is for an individual to be linked to only one household.
  - In addition, each farm can be viewed as a set of land parcels. In these Guidelines, the term “land parcel” generically indicates the elemental unit of land having a given area. Depending on how geographical information is prepared in a given country, this term may denote either:
    - A field that is a given portion of land and is operated by one farm in a uniform manner (e.g. has undergone the seeding of a certain type of plant, or a specific phytopharmacological treatment)
    - A definite point in a geographic grid with given latitudinal and longitudinal coordinates. In this case, the area is defined by adopting a convention, e.g. the area of a circle having a radius of two meters.

Thus, we can see that households and farms can be linked in various ways:

- One-to-one
- One-to-many/many-to-one
- Many-to-many.
The one-to-one option is most common for farms of the household sector, in which a farm is usually operated only by the components of a single household. This situation is illustrated in Figure 2.1 below, which depicts three households (H1, H2 and H3) and two farms (F1 and F2). Household H1 consists of three individuals: A (the head of household), B (who works in agriculture) and C (who does not work in agriculture). Individual A is also the agricultural holder of Farm F1. In this farm, individual B is also a worker. Therefore, Farm F1 is operated only by individuals of Household H1, namely individuals A and B. Accordingly, Farm F2 is operated by Household H2. Note that Farm F2 includes only individual E, the agricultural holder.

![Figure 2.1 – Example of one-to-one relation.](image)

In Figure 2.1 above, links are established between individuals to trace the same individual in the households and in the farms. However, links can also be drawn between clusters, to state in simple terms whether a household is linked to a farm and vice versa, regardless of how many links between individuals can be identified between the farm and the household.
Considering the links between the clusters, the situation portrayed in Figure 2.1 above can also be represented as follows (Figure 2.1.C):

![Figure 2.1.C – Situation exemplified in Figure 2.1 above, represented considering the links between clusters.](image)

The one-to-many and the many-to-one relations are described below in Figures 2.2a and 2.2b respectively. Figure 2.2 illustrates the case in which Household H1 operates two different farms: Individual A operates Farm F1, and Individual B is the agricultural holder of Farm F2.

![Figure 2.2.a – Example of a one-to-many relation.](image)
Figure 2.2.b below describes a common case among farms of the non-household sector, in which the farm includes individuals from several households.

![Diagram](image)

Figure 2.2.b – Example of a many-to-one relation.

Considering the links between clusters, the case is described in Figure 2.2.b.C below.

![Diagram](image)

Figure 2.2.b.C – Situation exemplified in Figure 2.2. above, represented considering the links between clusters.
Finally, the many-to-many situation, which is the most common example, is described in Figure 2.3.a below, in which Household H1 is related to Farms F1 and F2. Indeed, Individuals A and B work in Farm F1, while Individual C is the agricultural holder of Farm F2. Conversely, among Farm F1’s workers are individuals from Households H1 (A and B) and H2 (E). It may also be noted that Household H2 is related to Farms F1 (by means of individual E) and F2 (by means of individual D).

Figure 2.3.a – Example of a many-to-many relation.

Figure 2.3.b below presents the situation in which informal relationships can be observed and an individual may be linked to more than one farm, a situation that may arise in the agricultural sector. In particular, the figure illustrates the case in which Individual A is the agricultural holder of Farm F2 and works in Farm F2 (possibly for a limited number of days only).

Figure 2.3.b – Example of a more complex many-to-many relation.
It may be interesting to note that, in relation to the links between clusters, the cases of Figures 2.3.a and 2.3.b are described in the same way as in Figure 2.3.C, regardless of the multiple links of Individual A.

**Remark 2.1.** Another way to construct the links is to link households to farms only if a member of a given household is the farm’s agricultural holder (e.g. Figure 2.3.D). This is a common practice in developing countries, and was the rule followed by Burkina Faso in developing the frame for agricultural censuses. As stated in Example 5 below, this linking strategy may present some drawbacks in the estimation phase.
2.4 Observational methodology

The scheme is very simple and is based essentially on a two-step procedure.

- In the first step, data is collected in relation to the units of a given population, which may be either that of households or of farms. For specific units (e.g. a household), the data of two different types is collected:
  - the phenomena of interest relating to the unit (e.g. household consumption);
  - the links with the other population’s units.

- In the second step, the units of the populations that were not surveyed in the first step (e.g. farms) are interviewed using the links observed in that same step. All units having non-null links with the units observed in the first step are also observed. The collection of information gathers data of two different types:
  - the phenomena of interest relating to the unit (e.g. the number of poultry on the farm);
  - the links through which the unit can be reached from the units in the first step. This information is essential to compute the sampling weights (see Section 2.5 below).

There are two alternative possibilities:

- In the first step, the households are observed. Then, in the second step, the farms are observed.

- In the first step, the farms are observed. In the second step, the households are observed.

Examples

The examples below simulate the methodology to be applied when conducting a survey on the contribution of women’s employment to agriculture in developing countries. In this case, the access to and control of resources must be analysed, collecting data on the land owned or managed by women in male-headed farms, to evaluate the use of the inputs (improved seeds, machinery, and fertilizers) and the availability of credit and financial services.
For this reason, it is necessary to develop an integrated survey that can provide estimates of all inputs, by associating data at the household level on gender with data at the farm level on parcel/plot land tenure, and defining the links through the activity performed on the land as a female owner, manager or worker.

Depending on sample availability (household or farm), the relations concerning the statistical units observed are traced as follows, beginning with one of the two samples in the first step:

**Example 1: the household sample**

Consider a sample survey on the households selected by means of a standard stratified multistage sampling design. The Primary Sampling Units (PSUs) are stratified according to geographical criteria. The census’ Enumeration Areas (EAs) are the ultimate Sampling Units. All households belonging to a sampled EA are observed in the household sample.

Let us consider the situation exemplified in Figure 2.4 below, and suppose that the units H1 and H3 are selected in the sample (in two different EAs), with final inclusion probabilities (relating to stratification and multistage selection) respectively equal to:

- \( \pi_{H1} = 0.001 \)
- \( \pi_{H3} = 0.02 \)

meaning that Household H1 was selected from a set of 1000 units and Household H3 was selected from a set of 50 units.
Figure 2.4 – Example 1: Four households linked with two farms, of which the links between individuals or between clusters are represented.

**Step 1**  
The enumerator interviews the head of Household H1 and collects data on gender. Then, the enumerator collects data on the links with farms. These can be collected at:

- individual level (see the left-hand side of Figure 2.4 above), or
- cluster level (right-hand side).

For both types of observation, it is clear that the household is linked to Farms F1 and F2. The rule for observing the links (at individual or cluster level) is the same for all households in the sample.

The enumerator interviews the head of Household H3 and collects the data on gender. Then, the enumerator collects the data on the links with farms with the same method (i.e. at individual or at cluster level) adopted for Household H1. The enumerator verifies that the household is linked to Farm F1.
Step 2  The enumerator interviews the agricultural holder of Farm F1 and collects the data on the parcel/plot tenure. Then, the enumerator collects the data on the farm’s links. There are two possible alternatives:

- **If the links were collected at the individual level** (in Step 1), the enumerator will see that the farm features four individuals. Therefore, in this case, the value of the variable “links” is equal to 4.

- **If the links were collected at the cluster level**, the enumerator will see that the farm has individuals from three different households. Therefore, in this case the value of the variable “links” is equal to 3.

Next, the enumerator interviews the agricultural holder of Farm F2 and collects data on the parcel/plot tenure. The enumerator then collects the data on the farm’s links, following the same procedure adopted for Farm F1.

- **If the links were collected at the individual level**, the enumerator will see that the farm has two people. In this case, the value of the variable “links” is equal to 2.

- **If the links were collected at the cluster level**, the enumerator will see that the farm includes people from two different households. Thus, here too, the value of the variable “links” is equal to 2.

**Example 2: the farm sample**

Consider a sample survey on farms selected by means of a list frame with a stratified single-stage sampling design.

Let us consider the situation exemplified in Figure 2.5 below and suppose that Farm F1 is selected in the sample, with a final inclusion probability equal to

- \( \pi_{F1} = 0.04 \),

meaning that Farm F1 was selected from among other 25 units. Suppose also that in this case, the links are collected at the cluster level.
Figure 2.5 – Example 2: two farms linked with two households, with representation of the links between clusters.

**Step 1**
The enumerator interviews the agricultural holder of Farm F1 and collects data on parcel/plot tenure. Then, the enumerator collects data on the links with households and sees that the farm has individuals from Households H1 and H2.

**Step 2**
The enumerator collects data on gender from Household H1. Then, the enumerator collects data on the households’ links and sees that all the individuals involved in the households’ agricultural activities work on two farms. Therefore, in this case, the value of the variable “links” is equal to 2.

The enumerator collects data on gender from Household H2. Then, the enumerator collects data on the households’ potential links and sees that all individuals involved in the household’s agricultural activities work on one farm. Therefore, in this case the value of the variable “links” is equal to 1.

2.5 Estimation

Let $Y_H$ be the total of interest (e.g. the total number of women who perform agricultural activities in the country) referring to the population of the households, hereinafter denoted as $H$; and let $Y_F$ be the total of interest (e.g. the total land area in the country) referring to the population of the farms, hereinafter $F$. 
Let $S_H$ be the sample of observed households and $S_F$ the sample of observed farms.

The sample estimate $\hat{y}_H$ of the total $y_H$ is obtained by assigning to each household $h$ observed in $S_H$ a weight $w_h$ (generally greater than 1) that indicates how many households are represented by the observed household $h$. For example, a weight $w_h = 1,000$ indicates that the household $h$ represents itself, and that other 999 households are not included in the sample. The estimate $\hat{y}_H$ is then obtained as the weighted sum

$$\hat{y}_H = \sum_{h \in S_H} y_h w_h,$$

where $y_h$ is the value of the variable of interest in the household $h$ (e.g. the total number of women within the household who perform agricultural activities).

Likewise, the sample estimate $\hat{y}_F$ of the total $y_F$ is obtained by assigning, to each farm $f$ observed in $S_F$, a weight $w_f$ that indicates how many farms are represented by the observed farm $h$. The estimate $\hat{y}_F$ is then obtained as the weighted sum

$$\hat{y}_F = \sum_{f \in S_F} y_f w_f,$$

where $y_f$ is the value of the variable of interest in the farm $f$ (e.g. the total parcel/plot land tenure in the farm).

The weights $w_h$ and $w_f$ are computed differently depending on the statistical unit observed in the first step.

1. The weight of the unit observed in the first step of observation is computed in the traditional manner, as the reciprocal on the inclusion probability of the unit, and may be multiplied by a calibration/correction factor.

2. The weight of the unit observed in the second step of observation may be obtained by
   a. adding the weights (computed in the first step) of the units selected in the first step that are linked to the unit observed in the second step;
   b. dividing the above sum by the total number of the links of the (selected and non-selected) units.
Considering points (a) and (b) above, it can be seen that in one-to-one and one-to-
many relationships, the weight of the unit observed in the second step is always
equal to that of the unit observed in the first step.

**BOX 2.1**

**NOTE OF CAUTION**

This strategy produces unbiased estimates (FAO, 2014) for households and
farms if, and only if, the following two conditions hold:

- The sampling strategy for the units selected in the first step is roughly
  unbiased.
- Every unit in the population observed in the second step has at least one
  link with the units of the population observed in the first step.

**Remark 2.2.** Considering Box 2.1, the following question may arise: what would be
the best method for observing farms in the second step, if a country has several
existing household samples constructed with an unbiased sampling strategy? There
may be several reasonable choices. One possibility is to choose the household survey
with the closest objectives to those of the farm survey. Another alternative is to
choose the household survey that can maximize the size of the farm sample. This
would increase the accuracy of the farm survey.

**Households selected in the first step**

The weight of the household \( h \) is given by

\[
 w_h = \frac{1}{\pi_h} g_h,
\]

where \( \pi_h \) is the inclusion probability of the household \( h \) and \( g_h \equiv 1 \) is the calibration
 correction factor. To simplify the operation, we consider \( g_h = 1 \) in the sequel, and consequently:

\[
 w_h = \frac{1}{\pi_h}.
\]

Also, note that all individuals of the household \( h \) have the same weight \( w_h \).
The weight of the farm $f$ is obtained through the following procedure:

- First, all the weights of the sample households linked with the farm $f$ are summed up as
  $$\sum_{h \in S_H} w_h l_{h,f},$$
  where $l_{h,f}$ denotes the number of links between the household $h$ and the farm $f$. If the links between individuals are represented, then $l_{h,f}$ indicates the number of people of the household $h$ who work (or are more generally linked) with the farm $f$, if the links between clusters are represented, then $l_{h,f}$ is a binary variable equal to 1 if at least one individual of the farm $h$ is linked to the farm $f$.

- Then, the above sum is divided by the total number of the farm’s links, denoted by $L_f$. It is emphasized that this quantity is collected during the interview at the farm. Theoretically, $L_f$ is defined as
  $$L_f = \sum_{h \in H} l_{h,f}.$$
  Briefly, the weight $w_f$ is given by
  $$w_f = \frac{1}{L_f} \sum_{h \in S_H} w_h l_{h,f}.$$

**Example 3**

Let us consider again Example 1, illustrated in Section 2.4 above. The weights of Households H1 and H2 selected during the first step of the observation are equal to 1,000 and 50 respectively, and are represented as

- $w_{H1} = 1/\pi_{H1} = 1/0.001 = 1,000$ and
- $w_{H2} = 1/\pi_{H2} = 1/0.02 = 50$.

The weights of Farms F1 and F2, identified during the second step through the links with Households H1 and H2, are defined differently depending on how the links were observed.


Links observed at the individual level

Let us first consider Farm F1, and recall that there are two sample households H1 and H3, linked to the members of the farm by \( l_{H1,F1} = 2 \) and \( l_{H3,F1} = 1 \) individuals respectively. Consider also that the total number of links \( L_{F1} = 4 \). Therefore,

\[
w_{F1} = \frac{1}{L_{F1}} (2 w_{H1} + w_{H3}) = \frac{1}{4} (2000 + 50) = 512.
\]

Let us now consider Farm F2. The only sample household is H1, which links the farm F2 with a member (C), such that \( l_{H1,F2} = 1 \). Consider, also, that the total number of links \( L_{F2} = 2 \). Thus,

\[
w_{F2} = \frac{l_{H1,F2}}{L_{F2}} w_{H1} = \frac{1}{2} 1000 = 500.
\]

Links observed at the cluster level

The weights for Farms F1 and F2 are equal to 350 and 500 respectively, as

\[
w_{F1} = \frac{1}{L_{F1}} (w_{H1} + w_{H3}) = \frac{1}{3} (1000 + 50) = 350, \text{ and}
\]

\[
w_{F2} = \frac{1}{L_{F2}} w_{H1} = \frac{1}{2} 1000 = 500.
\]

To derive the above calculations, it is useful to recall that Household H2 was not part of the sample selected in the first step.

Farms selected in the first step

The weight of the farm \( f \) is given by

\[
w_f = \frac{1}{\pi_f g_f},
\]

where \( \pi_f \) is the inclusion probability of the farm \( f \), and \( g_f \approx 1 \) is the calibration correction factor. As for the households, let us consider that \( g_f = 1 \), such that

\[
w_f = \frac{1}{\pi_f}.
\]
The weight of the household $h$ is obtained by means of the following procedure:

- First, all the weights of the sample farms linked to the household $h$ are summed up as

$$
\sum_{f \in S_P} w_f l_{f,h},
$$
where $l_{f,h}$ denotes the number of links between the farm $f$ and the household $h$.

- Then, the above sum is divided by the total number of links of the household, $L_h$, where

$$
L_h = \sum_{f \in F} l_{f,h}.
$$

Briefly, the weight $w_h$ is given by

$$
w_h = \frac{1}{L_h} \sum_{f \in S_P} w_f l_{f,h}.
$$

**Example 4**

Let us consider again Example 2 of Section 2.4 above. The weight of Farm F1, selected during the first step of the observation, is equal to

$$
w_{F1} = \frac{1}{\pi_{F1}} = \frac{1}{0.04} = 25.
$$

The weights of Households H1 and H2, identified in the second step through the links with Farms F1 and F2, are equal to 12.5 and 25 respectively, as

$$
w_{H1} = \frac{1}{L_1} w_{F1} = \frac{1}{2} 25 = 12.5, \text{ and}
$$

$$
w_{H2} = \frac{1}{L_2} w_{F1} = \frac{1}{12} 25 = 25.
$$

To derive the above calculations, it is useful to recall that Farm F2 was not part of the sample selected during the first step.
**BOX 2.2**
**NOTE OF CAUTION**

If the links are defined properly, observing the household in the first step also enables the production of unbiased estimates for the population of farms. However, if the links are not defined properly, the estimates concerning the population of farms could be biased. This is shown in Example 5 below.

On the contrary, beginning the observation with the farms does not enable the production of unbiased estimates for the household population. Indeed, it is not possible to observe households that are not linked to the agricultural sector. This is clear from Figure 2.5; Household H3 cannot be observed with the mechanism described here.

Therefore, if the observation starts from the farms, unbiased estimates may be produced only for the population of the households related to agriculture.

---

**Example 5**

Let us consider the situation described in Remark 2.1 above (see Section 2.3), in which a household is linked to a farm only if the member of the household is the farm’s agricultural holder. Defining the links in this way makes it impossible to observe the farms that have an agricultural holder from a foreign country, as can occur with certain economic farms. This situation is described in Figure 2.6 below, in which it is supposed that Individual D, who is the agricultural holder of Farm F2, is a foreigner who is not permanently resident in the country. Obviously, given the definition of the links, it is impossible to observe Farm F2. On the other hand, if the links were defined at individual or cluster level, Farm F2 could be observed.

![Figure 2.6 – Example 5: links defined only through the agricultural holder.](image)
2.6 A note on the estimation of sampling errors

Suppose that a stratified multistage sampling is performed. Let

- \( r \) be a generic stratum (\( r = 1, \ldots, R \)) and
- \( i \) be the generic PSU.

Let us suppose that in the stratum \( r \), \( n_r \) PSUs have been sampled.

Approximating the variance with that of the sampling with replacement, the sampling variances of the estimates \( \hat{\gamma}_H \) and \( \hat{\gamma}_F \) can be estimated (Kish, 1966) as

\[
- \hat{\gamma}(\hat{\gamma}_H) = \sum_{r=1}^{R} \frac{n_r}{n_r - 1} \left[ \sum_{i=1}^{n_r} \hat{\gamma}_{Hi} - \left( \sum_{i=1}^{n_r} \hat{\gamma}_{Hi} / n_r \right) \right]^2,
\]

\[
- \hat{\gamma}(\hat{\gamma}_F) = \sum_{r=1}^{R} \frac{n_r}{n_r - 1} \left[ \sum_{i=1}^{n_r} \hat{\gamma}_{Fi} - \left( \sum_{i=1}^{n_r} \hat{\gamma}_{Fi} / n_r \right) \right]^2,
\]

where the totals at PSU level \( \hat{\gamma}_{Hi} \) and \( \hat{\gamma}_{Fi} \) are computed differently depending on the units (households or farms) selected in the first step.

Note that the above expressions coincide with the estimators of the variance derived by means of the Jackknife technique.

**Households selected in the first step**

The totals \( \hat{\gamma}_{Hi} \) and \( \hat{\gamma}_{Fi} \) are computed as follows:

\[
- \hat{\gamma}_{Hi} = \sum_{(h \in I) \cap (h \in \mathcal{S}_H)} y_h w_h
\]

\[
- \hat{\gamma}_{Fi} = \sum_{(h \in I) \cap (h \in \mathcal{S}_H)} w_h z_h,
\]

in which \( \sum_{(h \in I) \cap (h \in \mathcal{S}_H)} \) denotes the sum of the sampled households belonging to the \( t \)th PSU,

\[
\text{and } z_h = \frac{1}{L_f} \sum_{f \in S_f} w_f y_f l_{h,f}.
\]
Farms selected in the first step

The totals \( \hat{y}_{Hi} \) and \( \hat{y}_{fi} \) are computed as follows:

\[
\hat{y}_{Hi} = \sum_{(f \in i) \cap (h \in S_F)} z_f w_f
\]

\[
\hat{y}_{fi} = \sum_{(f \in i) \cap (f \in S_F)} w_f y_f ,
\]

in which \( \sum_{(f \in i) \cap (h \in S_F)} \) denotes the sum over the sampled farms belonging to the PSU,

and \( z_f = \frac{1}{L_h} \sum_{f \in S_F} w_f y_h l_{f,h} \).

2.7 Special case: subsampling the households of large farms

With regard to economic farms, the number of households to be interviewed is often too high. In this case, a simple solution is to subsample the households of the large farms.

Subsampling

The following procedure is proposed:

- **Establishing a threshold.** A threshold defining the maximum number of households to be observed per farm is established, having due consideration of the survey’s budget and relevant operational constraints. This number is denoted by \( m_{2H} \), e.g. \( m_{2H} = 10 \).

- **Identifying the number of households for each farm.** The households linked to a given farm are counted. This number of households is denoted with \( H_f \). For example, if the farm \( f \) includes 100 workers belonging to 15 different households, then \( H_f = 15 \).

- **Subsampling the farms.** If, for the farm \( f, H_f > m_{2H} \), then \( m_{2H} \) households are selected from the \( H_f \), using simple random sampling without replacement. Thus, in the farm \( f \), the sample \( \bar{s}_{2f} \) of \( m_{2H} \) households is selected. This represents a further stage of sampling in
which each household in the farm is selected, with a conditional inclusion probability equal to

\[ \bar{\pi}_{h|f} = \frac{m_{2H}}{H_f} \]

Let us note the following:

- The inclusion probability indicated above is denoted *conditional* since it depends on the fact that the farm \( f \) is selected in the sample of the first step.
- If all the households of a farm are observed (without subsampling), since \( H_f \leq \bar{\pi}_{2H} \), then

\[ \bar{\pi}_{h|f} = 1. \]

**Estimation**

As usual, the estimate \( \hat{Y}_H \) is then obtained as a weighted sum:

\[ \hat{Y}_H = \sum_{h \in S_H} y_h \cdot w_h, \]

where the weight of the household \( h \) is inflated by the factor \( 1/\bar{\pi}_{h|f} \):

\[ w_h = \frac{1}{L_f} \sum_{f \in S_f} w_f \cdot l_{f,h} \cdot \frac{1}{\bar{\pi}_{h|f}}. \]

**Example 7**

If the farm \( f \) comprises \( H_f = 20 \) households, and if the threshold is equal to \( m_{2H} = 10 \), then 10 households of the farm are selected and observed, each of which is included in the sample with a conditional inclusion probability of

\[ \bar{\pi}_{h|f} = \frac{m_{2H}}{H_f} = \frac{10}{20} = 0.5. \]
Suppose, also, that if there had been no subsampling, the sampling weight of the household $h$ would have been equal to 100. Then, having performed the subsampling procedure, the weight attached to the household $h$ is equal to

$$w_h = 100 \times \frac{1}{0.5} = 200.$$
3

Modules and operational rules for the application of indirect sampling in agricultural surveys

3.1 Overview

Developing countries have a particularly strong need to produce statistical data, develop their planning activities and measure the effectiveness of public interventions, especially in vulnerable areas. However, these countries do not always possess the resources required to build sampling frames for surveys that can provide reliable estimates at national level.

The rural sector is an especially vast and complex sector, such that it is not possible to collect all data on the sector during a general agricultural census. This is the case, for example, with the information on certain entities related to the agricultural sector, such as farmers’ organizations, agricultural cooperatives or agricultural training centres; for these, usually only information on their links with farms is collected. However, data for each of these structures are not collected, although requests to this effect arise often in developing countries.
This is also true for some information that is indirectly related to agriculture, such as its links with the sectors of education, health and roads.

For surveys in the rural sector, conventional sampling conducted through general agricultural censuses are not always suitable for collecting certain information, such as the production of rare crops. Indeed, developing countries often consider whether to promote certain crops that are high in value but are rarely cultivated by domestic producers. This type of crop requires specific sampling if high-quality estimates are to be produced.

Indirect sampling could help developing countries to obtain reliable data on the areas mentioned above. For each of these countries, it is necessary to diagnose information needs and existing data collections, and to analyse the possibility of using indirect sampling for integrated surveys in the rural sector while maintaining costs at a sustainable level.

This Chapter addresses the application of indirect sampling in the agricultural sector. It is divided into two main parts: (i) the observation of farms from a sample of households and (ii) the observation of households starting from farms. The Chapter will discuss the conceptual definition of the links between households and farms and the operational aspects of data collection; it will conclude with an illustration of some simulations.

### 3.2 The conceptual definition of links

#### 3.2.1 Issues

The conceptual definition of the links between statistical units is a very important step in the implementation of indirect sampling. Usually, several types of links are possible. The definition and selection of links must meet three important conditions, which are non-exhaustive:

- A *sine qua non* condition for the application of the methodology (non-zero link condition) is that each observed unit must be linked to at least one reference unit. This is necessary to achieve unbiased estimates;

- It is for the analysts to identify the links for which information can be collected as simply and reliably as possible. The pilot survey and pre-tests will be very useful for analysing the difficulties of collecting data on the links;
It is essential that the choice of links conform to the survey’s objectives for data analysis purposes (this aspect is discussed in further detail in Section 3.2.2 below, on the links between households and farms).

It is therefore essential that the identified links be defined rigorously and explained clearly to the investigators, for the data collection to yield as few non-answers concerning links as possible. Indeed, it should be recalled that each statistical unit linked will have to provide information on the total number of units of the reference population to which it is linked.

Some experts prefer to collect data on several types of links and then choose the most relevant \textit{ex post}: the drawback of this approach is an increased workload for investigators, which can negatively impact data quality. Box 3.1 below gives an overview of the issue of link choices.

**BOX 3.1**

**EXAMPLE – SURVEYING KINDERGARTEN CHILDREN WITHOUT A SAMPLING FRAME: A STUDY IN INDIRECT SAMPLING (Kiesl 2013)**

**Context**
One of the cohorts of the German National Educational Panel Study (NEPS) consists of a sample of kindergarten children. No nationwide frame of kindergartens in Germany was available for sample selection, unlike the situation for primary schools. Therefore, the indirect sampling method was used, and the links between kindergartens and primary schools were considered. Despite their spatial separation, kindergartens could be considered “linked” to primary schools, since every child who leaves a kindergarten eventually joins a particular primary school.

**Definition of links**
The links between a kindergarten b and a school a could be defined in two ways:

1) \( \text{Link} = 1 \), if at least one child moved from kindergarten b to school a during a given reference period. In this case, the total number of links of a kindergarten b is the number of schools that received children from it within the reference period;

2) \( \text{Link} = \text{number of children having moved from kindergarten b to school a during a particular reference period} \). In this case, the total number of links of a kindergarten b is the number of children that moved from that kindergarten b to any school within the reference period.

(cont’d)
Choice of links and justification

For every kindergarten b, the value of the total number of links had to be ascertained. In the case of link function (i), this required the number of primary schools to which the children who left kindergarten b during the reference period moved as first graders.

In the case of link function (ii), the number of children who left kindergarten b during the reference period and who joined any primary school had to be given. Pre-tests had shown that the latter information could be provided by kindergartens much more reliably. Kindergarten administrators are well aware of how many children have left, but usually do not know exactly which primary schools these children have joined, nor how many of them have done so. For these reasons, it was decided to use link function (ii) to construct the indirect sampling estimator.

3.2.2 Links between households and farms

Studies on the application of indirect sampling in agricultural surveys are rare in the literature. However, several links between a farm and a household can be examined. The following assumptions are possible, depending on the survey’s objectives and the socio-cultural circumstances of each country:

- **HP1**: a household is linked to a farm if the household owns all or part of this farm
- **HP2**: a household is linked to a farm if at least one household member works as a farmer on all or part of this farm, on behalf of the household
- **HP3**: a household is linked to a farm if at least one household member works on this farm (as a farmer or employee).

Let us conceptually analyse each of the above assumptions.

**a. Hypothesis HP1**

Hypothesis HP1 is very restrictive. It entails that for a given farm, the study must focus only on the households that own a property right over a portion of that farm. In this case, the households that practice farming through a lease or a loan of a part of the farm are not considered. Also, households having a member who works as a temporary or permanent employee on the farm are not taken into consideration.

In Figure 3.1.a below, which represents HP1, it is assumed that Farm F2 belongs to Household HP3. Thus, F2 is linked to HP3, although no member of HP3 works on F2.
Similarly, neither H1 nor HP2 are linked to F2, while some of their members work on the farm.

When choosing this link, it is necessary to ensure that every farm involved in the study belongs to a household. The number of these farms owned by the government, NGOs or other entities, must be negligible. Otherwise, this link is not suitable for the purposes of the indirect sampling method.

**b. Hypothesis HP2**

Hypothesis HP2 is less restrictive than Hypothesis HP1 because it takes into account all the households that are involved in farming but excludes employees. However, HP2 excludes some households considered by HP1: the households that own farms but that do not practice farming.

This is shown in Figure 3.1.b below: F2 is not linked to H3, although H3 is the owner of F2. Also, F2 is not linked to H1, while the individual C belonging to H1 works on F2 (as an employee).
Depending on the study’s objectives, before using this link, it should be ensured that there is only a negligible number of farms that are not operated by owners, or that are operated only by employees.

c. **Hypothesis HP3**
Finally, Hypothesis HP3 considers all households that participate in farms as farmers or employees, but does not take ownership into account.
To meet the requirements for the application of the indirect sampling methodology, it is essential to ensure that there is only a negligible number of farms that are not operated by their owners.

Other types of links can be defined, depending on the survey’s objectives. In Chapter 4, the operational constraints relating to each of the three links mentioned here will be discussed in further detail.

### 3.2.3 Level of link definition

As mentioned in Chapter 2, links can be defined at an individual level or at a household level. Each option affects the number of links between households and farms. The choice of level will affect data collection: it is important to choose the level that enables the most reliable collection of data concerning the links. In some situations, the information is more easily retrieved by considering the individual level; in others, collection at household level is preferable.

Figure 3.2 below compares the linkages at household and individual levels for Hypothesis HP3.

![Figure 3.2 – Comparison of linkages at household and individual levels.](image)
Part I - Modules and operational rules for observing farms starting from households

In Part I, it is assumed that a frame of agricultural households or a sample of households with the weight for each household is available. In general, in most countries the sampling frames of households are available with the population and housing census of the general agricultural census.

In this context, information on the farms used by the households is desired, but there is no sampling frame of farms. This Part will illustrate how indirect sampling can be used to collect data on a representative sample of farms and to calculate the weight of each farm using the households’ weights and the existing links between households and farms. It is important to note that even if a sampling frame of farms exists, indirect sampling can be used to collect data that integrates land and households.

1 Collecting data on the links

Collecting data on the links is a very important step in implementing the methodology, as the quality of these data is essential to achieve accurate estimates without bias. As indicated in the methodology, for each farm linked to households, the information sought during this collection is:

- the number and weights of the sampled households that are linked to the farm: the weights of the sampled households are known before the survey – if not, they must be estimated; the number of sampled households linked to the farm is estimated after the data collection (an example of an estimation procedure will be discussed in the section on simulations)
- the overall (nationwide) number of households related to the farm: this information must be collected during the survey.

Depending on the objectives of the study and the relevant socio-cultural context, it is necessary to define and identify the appropriate link between the statistical units (household and farm). Some examples of the possible links between these units and their influence on the results have already been seen above.
In practice, data on links can be collected through two separate questionnaires: these entail, on one hand, the identification and codification of the farms linked to the sampled households; and on the other, the information on the total number of links of the farms identified.

1.1 Identification and codification of farms

The purpose of this questionnaire is to identify all the farms that are linked to each sampled household and to codify them with a unique code. When several members of a given household work on several different farms, it is advisable to collect data from each member engaged in farming, and not only from one individual (e.g. the head of the household).

Each farm must be given a unique code, which must be independent of the household’s code. Indeed, it is important to consider that each farm observed becomes a new statistical unit for other types of data collection. A unique identifier for the farms also enables the identification of any duplicates and the estimation, for each farm, of the number of households within the sample and the total number of households that are linked to it. It is advisable to assign farm codes not when households are interviewed (since farm can be identified with reference to different households), but when the farms are being reached.

Concerning the codification of farms, it is important that different investigators use the same code for a given farm, in case they interview multiple households working on the same farm. Thus, a codification based on the farm’s geographical location will be more appropriate. If a complete list of all farms in the country exists, it is possible to create unique codes for farms in this list and then use these codes in the survey. Another solution could be the organization of a meeting with the enumerators after data collection with households, to assign a given farm to only one enumerator.

The questions to be asked when identifying the linked farms depend on the type of links adopted for the survey. For example, for the three types of links seen above, the active members of the household could be asked the identification questions set out in Table 3.1 below.
### Table 3.1 – Questions for the identification of farms.

<table>
<thead>
<tr>
<th>Links</th>
<th>Questions for identifying the farms linked to a household</th>
</tr>
</thead>
</table>
| HP1: A household is linked to a farm if the household owns all or a part of this farm. | Do you own farms that you acquired by inheritance, purchase or donation?  
If yes, how many farms do you have? Can you indicate to us the location of these farms for additional data collection? |
| HP2: A household is linked to a farm if at least one of its members works as a farmer on all or a part of this farm, on behalf of the household. | How many farms are you farming on behalf of your household?  
Can you tell us the location of these farms for additional data collection? |
| HP3: A household is linked to a farm if at least one of its members works on this farm (as a farmer or employee). | Have you worked on a farm last year (as a farmer or employee)?  
If yes, on how many farms?  
Can you tell us the location of these farms for additional data collection? |

### 1.2 Collecting data on the links of observed farms

After identifying the farms that are linked to a given household, the next step is to collect information on the total number of households linked to each farm. It is important for the investigator to find the individual who is best placed to answer this question. Indeed, the surveyed household may not necessarily be able to provide a reliable answer, especially if other households work on the same farm. For example, the investigator could ask the individual in charge of the farm for the exact number of households working on the farm. Here, again, the type of link used will determine the questions to be asked.

<table>
<thead>
<tr>
<th>Links</th>
<th>Questions for estimating the number of households linked to a farm</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP1: A household is linked to a farm if the household owns all or a part of this farm.</td>
<td>How many households own parcels on this farm?</td>
</tr>
<tr>
<td>HP2: A household is linked to a farm if at least one of its members works as a farmer on all or a part of this farm, on behalf of the household.</td>
<td>How many households have members who are working on this farm?</td>
</tr>
</tbody>
</table>
Links | Questions for estimating the number of households linked to a farm
---|---
part of this farm on behalf of the household. | N.B. This question could be difficult to answer if individual farmers are concerned. In these cases, it is preferable to consider the links at the individual level, and the question would be, simply, the following: How many farmers are there on this farm?

HP3: A household is linked to a farm if at least one of its members works on this farm (as a farmer or employee). | How many households have members working on this farm (as farmers or employees)?

N.B. This question too may be difficult to answer if dealing with individual farmers. In this case, it is preferable to consider the links at individual level. The question would then be: How many farmers and employees are there on this farm?

Table 3.2 – Questions for identifying the farms’ links.

1.3 Subsampling during the observation of farms

For practical and financial reasons, it is often difficult to collect data on all farmers of a given farm. This is true especially for large farms, which may have a great number of farmers.

One of the solutions proposed in literature is to perform a subsampling of farmers for data collection. Unbiased estimators are proposed to estimate indicators for the population of units observed (see Chapter 2). This procedure is developed in Lavallée (2007) under the term “two-stage indirect sampling”.

59
2 A practical example

We propose a practical example of the observation of farms from a household survey. This example is based on the socio-cultural context of Burkina Faso’s agricultural sector.

2.1 Description of the study

In Burkina Faso, irrigated crop production is performed mainly on a number of sites developed by the State, NGOs or other projects. On these sites, several farmers work individually or in groups to produce rice, corn, vegetables, etc. Suppose that a national study is to be conducted on certain characteristics of these sites: the area harvested, the crops, the number of farmers or the farmers’ incomes. It is also assumed that there is no reliable and updated sampling frame on these sites for sampling. In these cases, an indirect sampling will be performed using a sampling frame of households at national level.

2.2 Choice of the household frame

At national level, there are essentially two frames on households:

- The list of households from the General Census of Population and Housing (RGPH) and
- The database of agricultural households from the General Census of Agriculture (RGA).

Both frames can be used, but as mentioned in Chapter 2, it is preferable to use the most recent frame that is best suited to the study’s objectives. In this context, the RGA’s database appears to be the optimal choice: it is more recent than that of the RGPH and consists exclusively of households engaged in agriculture, and it is therefore likely to include members that are farmers on irrigated crop sites.

Instead, the RGPH frame includes all households, regardless of their activities. If a sample of households from the RGPH is used, the result will contain several households that do not have any connection with the irrigated sites. According to the indirect sampling methodology, this will not introduce any bias in our estimates, but, as mentioned in Chapter 2, it may negatively affect the accuracy of the estimations because a small number of irrigated crop sites would have to be investigated.
2.3 Selection of the household sample

After selecting the frame, there are two possible choices for the sample:

- Selection of a representative sample of households from the sampling frame
- Use of an existing sample selected from the frame and used in other surveys.

The second option appears best if other samples are already being used for agricultural surveys. This is for essentially two reasons:

- This sample will subsequently make more information available to analyse the data, because the survey will be linked to other ongoing surveys.
- It also has the advantage of working with a more up-to-date sample, reflecting the changes arising in the statistical units since the sampling frame was created.

In this example, Burkina Faso conducts permanent annual agricultural surveys, using a sample of households selected from the RGA sampling frame. Given the above considerations, his sample appears to be best suited to our study.

2.4 Choice of links

For this example, it is considered that the most suitable link between households and sites is the following: a household is linked to an irrigated production site if at least one of its members works on this site as a farmer or employee. The reasons for this choice are described below.

First, the link is defined at individual and not household level: it is known that on the sites, parcels are provided to farmers as individuals, and members of the same household can work on different farms. Generally, the head of a site’s management unit is aware of the number of farmers on the site, but not of the number of households whose members work there. Defining the link at household level would make it difficult to collect data on the overall number of households linked to each site under observation.

Then, a household is linked to a site when one of its members works as a farmer or employee on the site: it is known that, given the limited number of developed sites, several producers cannot obtain a parcel for irrigated production on a developed site.
Thus, for many households, the links to irrigated sites are limited to the work of some of their members as permanent or temporary employees on these sites. The inclusion of employees and not only farmers in the link will enable a greater number of sites to be covered in our sample.

2.5 Collecting data on the links

To collect data on the links, two main questionnaires are proposed, one on household links and one on site links.

2.5.1 Questionnaire on household links

For each household, information must be collected on the number of sites where household members work as farmers or employees and to identify each site.

It is essential that each of the household’s active members answer the questionnaire, to ensure availability of the greatest amount of reliable information possible.

<table>
<thead>
<tr>
<th>Household Code: ..........................</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active member code: ........................</td>
</tr>
<tr>
<td>Were you a farmer on an irrigated site last year? ..........................</td>
</tr>
<tr>
<td>If yes, list the sites concerned:</td>
</tr>
<tr>
<td>• Name of Site 1.............................</td>
</tr>
<tr>
<td>• Name of Site 2.............................</td>
</tr>
<tr>
<td>• Name of Site 3.............................</td>
</tr>
<tr>
<td>• Name of Site 4.............................</td>
</tr>
<tr>
<td>Were you an employee on an irrigated site last year? ...................</td>
</tr>
<tr>
<td>If yes, list the sites concerned:</td>
</tr>
<tr>
<td>• Name of Site 1.............................</td>
</tr>
<tr>
<td>• Name of Site 2.............................</td>
</tr>
<tr>
<td>• Name of Site 3.............................</td>
</tr>
<tr>
<td>• Name of Site 4.............................</td>
</tr>
</tbody>
</table>

Table 3.3.a – Questionnaire to ascertain household links.
**Summary of sites linked to the household**

This section must be completed by the investigator, who must summarize which sites are linked to the household and assign a code to each. For each site, the investigator must choose a household member as a respondent, to assist him in identifying the site and collecting data on it.

<table>
<thead>
<tr>
<th>1.</th>
<th>Site Name</th>
<th>Member Code</th>
<th>Respondent</th>
<th>Site Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>...........</td>
<td>...........</td>
<td></td>
<td>...........</td>
</tr>
<tr>
<td>2.</td>
<td>...........</td>
<td>...........</td>
<td></td>
<td>...........</td>
</tr>
<tr>
<td>3.</td>
<td>...........</td>
<td>...........</td>
<td></td>
<td>...........</td>
</tr>
<tr>
<td>4.</td>
<td>...........</td>
<td>...........</td>
<td></td>
<td>...........</td>
</tr>
<tr>
<td>5.</td>
<td>...........</td>
<td>...........</td>
<td></td>
<td>...........</td>
</tr>
<tr>
<td>6.</td>
<td>...........</td>
<td>...........</td>
<td></td>
<td>...........</td>
</tr>
</tbody>
</table>

Table 3.3.b – Questionnaire to ascertain household links (Investigator section).

### 2.5.2. Site links

The essential information to be collected on each site is the total number of its links to households. Farmers and employees from other countries should be excluded. The questionnaire designed should incorporate information on the site code, which must be the same as that of the questionnaire on the households’ links. As specified above, the latter code should be the same for all the investigators who interview households that are linked to the same site. For example, geo-referential coordinates are suitable as site codes.

<table>
<thead>
<tr>
<th>Site Name:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site Code:</td>
</tr>
<tr>
<td>Name of the Respondent on the Site:</td>
</tr>
</tbody>
</table>

**Questions for the Respondent**

- In total, how many farmers (residents in the country) are working on this site? .................
- Can you provide an estimate of the total number of employees (excluding farmers) working on the site? .................

Table 3.4 – Questionnaire for the sites’ potential links.
Respondents will almost certainly be unable to provide a reliable answer to the latter question, especially for large sites. For this reason, it is preferable to select a sample of the site’s farmers, to estimate the average number of employees per farmer and thus the total number of employees working on the site.

3 Simulation of the estimation of farms’ weights

The theoretical basis for calculating the weight of farms is explained in Chapter 2. Here, a simulation of the calculation of farm weights through the GWSM is proposed, using sample data collected from the example illustrated above.

3.1 Recalling the method of weight estimation using the GWSM

As shown in Chapter 2 above, the process for calculating the weight is the following:
3.2 A numerical example

In this example, the weight of a farm that is linked to three sampled households and five other non-sampled households is calculated.

<table>
<thead>
<tr>
<th>HH code</th>
<th>Member code</th>
<th>Farm code</th>
<th>HH weight</th>
<th>Farm Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>563</td>
<td>21</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>563</td>
<td>47</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>563</td>
<td>21</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>563</td>
<td>82</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>563</td>
<td>47</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1365</td>
<td>21</td>
</tr>
</tbody>
</table>

Figure 3.4 – Numerical example of weight calculation.
3.3 Excel simulation

We create a data simulation with Excel software to provide an example of a data processing step to be performed after the collection phase. We consider a simple example, with two households linked to three farms. The households’ links are represented in Figure 3.5 below. It must be noted that the figure does not show the overall situation of these households and farms, because the members of households that are not linked to farms are not represented; in addition, other households that may be linked to the three farms are not represented.

Figure 3.5 – Illustration of the simulation example.

<table>
<thead>
<tr>
<th>Legend</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HH code</td>
<td>Household code</td>
</tr>
<tr>
<td>Member code</td>
<td>Household member code</td>
</tr>
<tr>
<td>Farm code</td>
<td>Code of the farm linked to household member</td>
</tr>
<tr>
<td>HH weight</td>
<td>Household weight</td>
</tr>
<tr>
<td>Farm links</td>
<td>Farm links (total individuals working in the farm)</td>
</tr>
</tbody>
</table>

Figure 3.6 – Simulation data structure and legend.
In an Excel workbook, the data for these two households may emerge as described below. This is the information available after collection and data entry. The method applied to process this data to calculate the weight of the parcels is illustrated below.

Step 1: Sort the database by farm code

![Figure 3.7.a – Step 1: weight calculation.]

<table>
<thead>
<tr>
<th>HH code</th>
<th>Member code</th>
<th>Farm code</th>
<th>HH weight</th>
<th>Farm Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>563</td>
<td>21</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>563</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1365</td>
<td>21</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>563</td>
<td>82</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>563</td>
<td>47</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>563</td>
<td>47</td>
</tr>
</tbody>
</table>

Step 2: Calculate the farm’s initial weight

**Reminder:** The farm’s initial weight is the sum of the weights of all the sampled households linked to it.

![Figure 3.7.b – Step 2: weight calculation.]

<table>
<thead>
<tr>
<th>HH code</th>
<th>Member code</th>
<th>Farm code</th>
<th>HH weight</th>
<th>Farm Links</th>
<th>Farm initial weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>563</td>
<td>21</td>
<td>2491</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>563</td>
<td>21</td>
<td>2491</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1365</td>
<td>21</td>
<td>2491</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>563</td>
<td>82</td>
<td>563</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>563</td>
<td>47</td>
<td>1126</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>563</td>
<td>47</td>
<td>1126</td>
</tr>
</tbody>
</table>
Step 3: Calculate the farm’s final weight

**Reminder:** The farm’s final weight is the division of its initial weight by its potential links.

<table>
<thead>
<tr>
<th>HH code</th>
<th>Member code</th>
<th>Farm code</th>
<th>HH weight</th>
<th>Farm Links</th>
<th>Farm initial weight</th>
<th>Farm final weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>563</td>
<td>21</td>
<td>2491</td>
<td>118.62</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>563</td>
<td>21</td>
<td>2491</td>
<td>118.62</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1365</td>
<td>21</td>
<td>2491</td>
<td>118.62</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>563</td>
<td>82</td>
<td>563</td>
<td>6.87</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>563</td>
<td>47</td>
<td>1126</td>
<td>23.96</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>563</td>
<td>47</td>
<td>1126</td>
<td>23.96</td>
</tr>
</tbody>
</table>

Figure 3.7.c – Step 3: weight calculation.
Part II – Observing households starting from a survey on farms

Part II will examine the case of data collection on households from a sample of farms. In practice, this situation is not as common as that seen in Part I above (observation of farms from a sample of households), but it can arise in practice.

For example, consider the situation of a country that has a sound information system for farms, and has a sampling frame on all farms or on those that produce a specific crop or use a particular agricultural technique (e.g. mango farms, irrigated areas using drip, etc.). In this case, it is possible to attempt to collect information on all households associated with these farms. Indirect sampling will be very useful in this respect, especially when there is no frame of households involved in the study. Therefore, it is possible to use a sample of farms for information on all the households concerned.

1 Definition of links

We have already discussed the possible links between a household and a farm on the basis of the three assumptions below, which may be traced depending on the survey’s objectives and the socio-cultural situations of each country:

- HP1: a household is linked to a farm if the household owns all or part of this farm.
- HP2: a household is linked to a farm if at least one household member works as a farmer on all or part of this farm on behalf of the household.
- HP3: a household is linked to a farm if at least one household member works on this farm (as a farmer or employee).

Let us consider the implications of each hypothesis in the case of a household survey performed through an observation of farms:

- Hypothesis HP1 takes into account the farms that are wholly or partially owned by households. If, within the study, several farms belong to the State, NGOs, producer associations or other entities, these farms will have no link to households. This assumption is well-suited to a study that focuses only on landowning households. The advantage of this hypothesis is that the data on the links will be easy
to collect, because the list of households having a right of property in a particular farm is easy to establish.

- Hypothesis HP2 is used to collect data on all households whose members are farmers, but excludes households that have owners but no farmers.
- Hypothesis HP3 concerns all farmers and their employees. It enables a large number of households to be studied, but its disadvantage lies in the fact that information on the links is more difficult to collect: indeed, for a given farm, it is necessary to collect information on the number of households whose members are farmers or employees within it.

2 Data collection on the links

Data collection on the links will enable two important sets of information to be gathered: the number of each farm’s links to households and the total number of links of each household observed.

2.1 Number of links of each farm

For each farm sampled, the aim is to ascertain the number of households to which it is linked. Depending on the level of link definition (individual or household) and the type of link, different questions can be asked to collect this information (see Table 3.5 below).

<table>
<thead>
<tr>
<th>Links</th>
<th>Question to identify the number of links of each farm</th>
</tr>
</thead>
</table>
| HP1: A household is linked to a farm if the household owns all or a part of this farm | – How many households have a property right over this farm? (household-level link)  
– How many farmers have parcels on this farm? (individual-level link) |
| HP2: A household is linked to a farm if at least one of its members works as a farmer on all or a part of this farm on behalf of the household | – How many households have members who are farmers in this farm? (household-level link)  
– How many farmers are there in this farm? (individual-level link) |
| HP3: A household is linked to a farm if at least one of its members works on this farm (as a farmer or employee) | – How many households have members who are farmers or employees in this farm? (household-level link)  
– How many farmers and employees are there on this farm? (individual-level link) |

Table 3.5 – Questions on farms’ links for each hypothesis.
As emphasized in Part I above, the units observed (linked to the sampled units) must be codified, with the assignment of a unique code to each unit; all investigators should use the same codification method. This is necessary to estimate the total number of the links of each observed household with the sampled farms, which will be calculated after the data collection.

### 2.2 Subsampling of the households observed

In theory, all the households linked to a given farm should be interviewed. The main problem that may arise is that some farms may be linked to a large number of households (in the case of large farms, a number that can often reach several hundreds). It would be difficult to investigate each one of these households.

In literature, a suggested solution is to select a sample of the linked households for observation. Unbiased estimators are proposed for this approach (Kiesl, 2013).

### 2.3 The total number of links of observed households

After selecting the households to be observed, the next step is the collection of the total number of links of each household with all farms at national level. At this stage too, the questions must be accurate and may vary depending on the type of link selected.

<table>
<thead>
<tr>
<th>Links</th>
<th>Questions to identify the number of each household link observed</th>
</tr>
</thead>
</table>
| HP1: A household is linked to a farm if the household owns all or a part of this farm | – In how many farms do members of your household own parcels? (household-level link)  
– How many parcels on farms do the members of your household have? (individual-level link) |
| HP2: A household is linked to a farm if at least one of its members works as a farmer on all or a part of this farm on behalf of the household | – In how many farms are members of your household farmers? (household-level link)  
– Each member of your household is a farmer on how many farm parcels? (individual-level link) |
| HP3: A household is linked to a farm if at least one of its members works on this farm (as a farmer or as an employee). | – In how many farms are members of your household farmers or employees? (household-level link)  
– Each member of your household is a farmer or an employee on how many farm parcels? (individual-level link) |

Table 3.6 – Questions to ascertain households’ potential links under different hypotheses.
3 A practical example

Consider the case of a country that wishes to collect information on the households engaged in farming on orange farms at the national level. Assume that orange farming is a very important sector in the agricultural economy of that country; thus, information on the characteristics of the households involved is necessary to enable more effective government intervention in the sector.

It is also assumed that the country has a sampling frame on orange farms that can be used to rigorously select a sample of farms, from which to observe households through indirect sampling.

3.1 Choice of link

For this example, the choice of link is intrinsically imposed by the study’s objective, which focuses upon the households working on orange farms. The link to be selected corresponds to Hypothesis HP2: *a household is linked to a farm if at least one of its members works as a farmer on all or a part of this farm on behalf of the household.*

For this type of farm, it is assumed that it is generally easier to discover the number of households than the number of individual farmers. Indeed, these types of farm are usually owned by several households (individual farmers are rare); therefore, it will be easier for the individual in charge of the farm to provide the number of working households.

3.2 Data collection on links

As in the case of Part I above, data on links can be collected by means of two main questionnaires: one on the farms’ links and another on the households’ links. We also assume that several households usually work on this type of farm; thus, it will be useful to create a sub-sample of households. Assuming that the households linked to a given farm are relatively homogeneous with respect to the information sought in the study, the sampling will be performed as follows:

- if a sampled farm is linked to more than 10 households, a simple random sampling of 10 households will be performed;
- if the total number of households linked to a sampled farm is lower than 10, all the households linked to the farm will be interviewed.
**Questionnaire to ascertain the farms’ links**

This questionnaire can be used to collect data on the total number of households linked to each farm and the identification codes of the linked households that were sampled.

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Full name of head of household</th>
<th>Number of working members</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.7.a – Questionnaire to ascertain the farms’ links.

**Identification of sampled linked households**

As mentioned earlier, it is useful to codify the linked households through a unified approach, which allows different investigators to use the same code for the same household. For this purpose, the number of the head of the household’s national identity card will be chosen as an identifier of the household. The interviewer should also leave instructions to provide the same number to any other investigator who may investigate the household.
Questionnaire on the households’ links

For each household linked, information must be collected on the total number of farms to which the household is linked.

<table>
<thead>
<tr>
<th>Farm code</th>
<th>Serial number of the household in Farm</th>
<th>Household identifier</th>
<th>In how many farms are members of your household farmers?</th>
</tr>
</thead>
</table>

Table 3.8 – Questionnaire to ascertain households’ potential links.

3.3. Weight calculation

According to the GWSM, with a subsampling of the units observed, the process for calculating the weight of each household is the following:

Figure 3.8 – Illustration of the weight calculation.
Numerical example

**Figure 3.9 – Numerical illustration of the weight calculation.**
3.4 Simulation with Excel data

During this stage, the weight calculation using the GWSM is simulated with a subsampling of the units to be observed, as described in the example above. In this simulation, three sampled farms and four households linked to these farms will be considered. These households were sampled for data collection. The essential data processing steps to calculate the weight of households are described.

The links between the farms and households in the simulation are shown in Figure 3.10 below.

![Figure 3.10 – Illustration of the simulation example.](image)
The raw data collected provide the information set out in the datasheet below.

<table>
<thead>
<tr>
<th>Farm code</th>
<th>Farm weight</th>
<th>Nb of HH on the farm</th>
<th>Nb of obs HH selected</th>
<th>Simulation HH code</th>
<th>HH Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>53</td>
<td>10</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>53</td>
<td>10</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

**Legend**

- **Nb of HH on the farm**: Total number of households working in the farm
- **Nb of obs HH selected**: Number of observed households sample
- **Simulation HH code**: Code of sampled observed household considered in the present simulator
- **HH Links**: Total number of the household's links
- **HH inv selection prob.**: Household's inverse selection probability
- **HH initial weight**: Household's initial weight
- **HH final weight**: Household's final weight

**Figure 3.11** – Data structure and legend of the simulation example.

**Step 1: Sort database by household code**

<table>
<thead>
<tr>
<th>Farm code</th>
<th>Farm weight</th>
<th>Nb of HH on the farm</th>
<th>Nb of obs HH selected</th>
<th>Simulation HH code</th>
<th>HH Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>53</td>
<td>10</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>53</td>
<td>10</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

**Figure 3.11.a** – Step 1: Calculation of households' weights.
### Step 2: Calculation of the households’ inverse selection probability

<table>
<thead>
<tr>
<th>Farm code</th>
<th>Farm weight</th>
<th>Nb of HH on the farm</th>
<th>Nb of obs HH selected</th>
<th>Simulation HH code</th>
<th>HH Links</th>
<th>HH inv selection prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>1</td>
<td>5</td>
<td>1.8</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>53</td>
<td>10</td>
<td>1</td>
<td>5</td>
<td>5.3</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>1.8</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>1.8</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>3</td>
<td>11</td>
<td>1.8</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>4</td>
<td>7</td>
<td>1.8</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>53</td>
<td>10</td>
<td>4</td>
<td>7</td>
<td>5.3</td>
</tr>
</tbody>
</table>

*Figure 3.11.b – Step 2: Calculation of households’ weights.*

### Step 3: Calculation of households’ initial weights

<table>
<thead>
<tr>
<th>Farm code</th>
<th>Farm weight</th>
<th>Nb of HH on the farm</th>
<th>Nb of obs HH selected</th>
<th>Simulation HH code</th>
<th>HH Links</th>
<th>HH inv selection prob.</th>
<th>HH initial weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>1</td>
<td>5</td>
<td>1.8</td>
<td>200.7</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>53</td>
<td>10</td>
<td>1</td>
<td>5</td>
<td>5.3</td>
<td>200.7</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>1.8</td>
<td>76.6</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>4</td>
<td>7</td>
<td>1.8</td>
<td>219.7</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>76.6</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>53</td>
<td>10</td>
<td>4</td>
<td>7</td>
<td>5.3</td>
<td>219.7</td>
</tr>
</tbody>
</table>

*Figure 3.11.c – Calculation of households’ weights.*
### Step 4: Calculation of households’ final weights

#### Figure 3.11.d – Step 4: Calculation of households’ weights.

<table>
<thead>
<tr>
<th>Farm code</th>
<th>Farm weight</th>
<th>Nb of HH on the farm</th>
<th>Nb of obs</th>
<th>HH selected</th>
<th>Simulation HH code</th>
<th>HH Links</th>
<th>HH inv selection prob.</th>
<th>HH initial weight</th>
<th>HH final weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>1</td>
<td>5</td>
<td>1.8</td>
<td>200.7</td>
<td>40.1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>76.6</td>
<td>38.3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>53</td>
<td>10</td>
<td>3</td>
<td>11</td>
<td>1.8</td>
<td>57.6</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>4</td>
<td>7</td>
<td>1.8</td>
<td>219.7</td>
<td>31.4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>219.7</td>
<td>31.4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>53</td>
<td>10</td>
<td>4</td>
<td>7</td>
<td>5.3</td>
<td>219.7</td>
<td>31.4</td>
<td></td>
</tr>
</tbody>
</table>

Note: The calculation of HH final weight is given by the formula: 

$$HH	ext{ final weight} = \frac{HH	ext{ initial weight}}{HH	ext{ selection prob.}}$$
Changes in statistical units

4.1 Introduction

In several developing countries, annual surveys are conducted in the agricultural sector to collect information essential to the population’s food security and to assess the agricultural sector’s performance. These surveys are usually conducted with a panel of agricultural holdings selected from the general agricultural census data. The use of panels certainly presents several advantages in statistical and operational terms, but many technical constraints in its implementation may arise.

Among these challenges is the fact that the statistical units used for these surveys may experience significant changes (disappearance, division, or merger) from one year to another due to endogenous or exogenous events. Phenomena occurring in the population may impair the sample’s quality. These changes adversely affect the quality of panels because they directly influence sample size and the weight of the statistical units. However, these panels are mainly used in developing countries for cross-sectional analysis, unlike other panels, the main objective of which is inter-temporal (longitudinal) analysis. Indeed, panel surveys are generally used for these countries because they do not require further annual expenditures to be incurred for a new sample selection.

This Chapter will focus on the changes to statistical units and the solutions that can be adopted to ensure the reliability of agricultural surveys’ data.
4.2 Analysis of changes (households and farms)

The statistical units of panel surveys can experience three types of change: disappearance, division or merger. To these, certain population changes that can affect sample quality must be added. These changes have a direct effect on cross-sectional analysis and can significantly affect the sample’s representativeness and the weight of statistical units. However, it must be noted that when the changes are random, they have no effect on the sample’s representativeness, and therefore no correction for cross-sectional estimates is necessary. This is the case with natural phenomena such as births and deaths.

4.2.1 Disappearance

The statistical units of a longitudinal survey may disappear from the geographical area covered by the study from one year to the next. For households, the disappearance is usually related to death or migration, and generally concerns smaller households consisting of one or two members. Cases where multiple household members disappear for a given area are quite rare, although are possible due to conflicts, migration or natural disasters. The phenomenon of disappearance also accounts for the absolute refusal of a statistical unit to collaborate in data collection, due for example to fatigue or other reasons. For farms, the loss may be related to farmers’ abandonment of land (e.g. because of a set-aside), conflicts or a natural phenomenon (e.g. degradation caused by drought).

4.2.2 Fusion

A household may welcome members of other households due to marriage, employment or migration. It is even possible for two households of the sample to merge completely and form a single household. This phenomenon can also be observed in farms. From one year to the next, one farm of the sample can merge to become part or all of another farm.

4.2.3 Division

A household may lose many members for reasons such as marriage, migration or death. Farms may also suffer a loss of area due to degradation or land conflicts.
4.2.4 Population change

Certain significant phenomena may cause structural changes in the population and affect the sample’s representativeness. This may be due, for example, to massive migration into the country during a given year, with the consequence that new households enter the country. Major public interventions in the agricultural sector during a given year may significantly modify the agricultural economy of the country, so that the panel sample is not representative. For example, if the government decides to support a particular crop, the proportion of households producing that crop may change significantly. Certain exogenous factors, such as rising international prices, may also considerably change the structure of the agricultural activities performed. It must be noted that structural changes in the economy usually occur after several years. This is why regular renewal of the panel (every 5 or 10 years) is recommended to ensure sample representativeness.

4.3 The main solutions for correcting sample changes

4.3.1 Additional samples

Due to population movements, new statistical units may appear in the population (of households or farms) over two years. If the proportion of new units is significant, an additional sample of this new population must be selected so that a representative sample for cross-sectional estimates may be obtained. During this operation, the population of new units can be considered a stratum.

Example

Suppose that due to a phenomenon linked to unemployment in the urban area or to increased farming profitability, several non-farm households decide to invest in agriculture. Thus, the population of agricultural households will increase dramatically with the entry of these new households. The panel’s initial sample will not be representative of the current agricultural population. To resolve this issue, for data collection, a representative sample of new households can be taken and added to the initial sample.

4.3.2 Oversampling

Another solution to the problem of the disappearance of statistical units is oversampling, i.e. increasing the sample size to anticipate the loss of statistical units. This helps to maintain sample accuracy, but does not prevent bias.
**Example**

When designing a given panel, suppose that the sample size required for data collection is estimated at 2,500 households. Previous investigations have shown that overall, 7% of households disappeared during the period of data collection with the panel. To take this phenomenon into account, the new size of our sample will be $2,500 \times (1 + 7\%) = 2,675$ households.

### 4.3.3 Post-stratification

The post-stratification technique is another solution used in data processing to avoid bias relating to sample representativeness. Indeed, the structural changes in the population can require the statistician to adjust the weights of the strata or create new strata.

**Example**

Suppose that in the initial sample, the following three strata were chosen:

- Stratum 1: food crop farmers
- Stratum 2: cash crop farmers
- Stratum 3: vegetable crop farmers

The sample was drawn during Year 1, considering these strata and their proportions given in Table 4.1 below. Suppose that during Year 3, due to food insecurity problems in Year 2, the government decided to provide special support to producers of food and vegetable crops (strata 1 and 3) with large subsidies on inputs. Thus, many producers abandoned the cultivation of cash crops in favour of subsidized crops. This will lead to a change in the structure of the strata (see Table 4.1 below). Therefore, the sample does not accurately represent the population during Year 3, and the estimates will certainly be imprecise. For example, if the total number of producers of vegetable crops must be estimated, the sample will provide results that do not conform to reality. Post-stratification adjustments calculate post-stratification coefficients to adjust the weight of each statistical unit on the basis of its stratum population values.
### Table 4.1 – Example of post-stratification application.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Proportion Year 1</th>
<th>Proportion Year 3</th>
<th>Adjustment factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strata 1</td>
<td>0.54</td>
<td>0.61</td>
<td>1.13</td>
</tr>
<tr>
<td>Strata 2</td>
<td>0.31</td>
<td>0.16</td>
<td>0.52</td>
</tr>
<tr>
<td>Strata 3</td>
<td>0.15</td>
<td>0.23</td>
<td>1.53</td>
</tr>
</tbody>
</table>

**4.3.4 Tracking**

As mentioned above, changes in statistical units adversely affect their representativeness and make estimates less precise. These changes must be corrected to maintain the integrity of the units. When a part of a unit does not exist at the time of collection, this part will have to be tracked for data collection, especially if the absence is not random. In the case of households, if, for example, one or more household members are absent or are no longer considered its members at the time of collection, they should be sought for the purposes of data collection. Similarly, if a portion of a farm changes ownership e.g. due to a land conflict, arrangements should be made with the new owner to collect data on this part. The example below describes the process, in Burkina Faso’s agricultural surveys, of tracking household members when the statistical unit is divided.

**Example: Tracking process in Burkina Faso**

In Burkina Faso’s agricultural statistical system, the primary statistical unit is the village and secondary statistical unit is the household. Let us assume that in a given village, a household is divided into two groups: Group A and Group B. Table 4.2 below lists the instructions given to investigators depending on the situation of the two groups.
### 4.4 Changes in households and weight calculation

In several countries (and in the case of most developing countries), statistical surveys use households as statistical units, and the data are collected at the household level. In this case, the solutions below will usually be sufficient to solve problems relating to changes in the composition of households for cross-sectional estimates in a given year.

The system may need to examine the activities, pursued by individual household members, that are relevant to agricultural statistics. In this case, each member becomes a statistical unit and the weight of a member is equal to that of his household. Indeed, all members of a sampled household are usually taken into account in the survey, and each year, data on all individual members of the household panel must be collected for cross-sectional estimates. In each wave of the panel, all households with at least one of these individuals as a member must be taken into consideration. For this type of system, changes in statistical units can lead to situations in which it is difficult or impossible to calculate the weights of certain units by means of the conventional approach of using the inverse of the inclusion probabilities. For example, several members of different households may combine to form a new household that also includes members of non-sampled households. The weights of all the members of the new household thus become difficult to compute.

<table>
<thead>
<tr>
<th>Context</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A is installed independently in the village, and Group B is also installed independently in the village; both are independent entities.</td>
<td>The investigator must survey Groups A and B as part of the same household.</td>
</tr>
<tr>
<td>Group A is installed independently in the village and Group B joins another household in the village.</td>
<td>The investigator must survey only Group A and ignore Group B.</td>
</tr>
<tr>
<td>Group A is installed independently in the village and Group B migrates away from the village.</td>
<td>The investigator will survey only Group A.</td>
</tr>
<tr>
<td>Both groups A and B are out of the village, or each merge to another household of the village.</td>
<td>The investigator ignores the household.</td>
</tr>
</tbody>
</table>

*Table 4.2 – Example of household tracking in Burkina Faso.*
The literature also suggests other methods for calculating the weights. Basically, there are weight-sharing methods, which use the weight of the households of the first wave of the panel (Brick and Kalton, 1994; Schonlau et al., 2013; Lavallée, 2007), and methods based on a model of household inclusion probabilities (Schonlau et al., 2013). In this Section, methods for sharing weight are discussed, with particular emphasis on the GWSM described in Chapter 2.

These methods concern all households of which at least one member was a member of a household sampled during the first wave of the panel. Some target households may have members who were in the original sample panel and other members who were not. Weight-sharing methods only enable the calculation of the weights of the households containing members that were within the original sample. For agricultural surveys, these weights are usually sufficient for a cross-sectional estimate of key agricultural sector indicators.

### 4.4.1 Classic methods of weight-sharing

Basically, there are two approaches: the equal household weighting scheme, and the equal person weighting scheme.

#### 4.4.1.1 Equal household weighting scheme

Consider a household $h$ with $m_h$ members. For every household $j$ that existed during the selection of the initial sample (first wave), and at least one member of which was within household $h$, $w_{hj}$ is considered equal to the weight of $j$ if $j$ was selected in the initial sample of the panel and $w_{hj} = 0$ otherwise. The weight $w_h$ of $h$ can be written as

$$w_h = \sum_{j=1}^{m_h} \alpha_{hj} w_{hj},$$

with

$$\sum_{j=1}^{m_h} \alpha_{hj} = 1,$$

in which $\beta_h$ is the number of households in the initial sample that feature a member in household $h$. 

86
Huang (1984) proposes the use of $\alpha_{hj} = \frac{1}{\beta_h}$. This gives

$$w_h = \frac{1}{\beta_h} \sum_{j=1}^{\beta_h} w_{hj}.$$  

### 4.4.1.2 Equal person weighting scheme

This approach is similar to the one seen above. In this case, for any individual member $k$ of a household $j$ that existed during the selection of the initial sample (Year 1) and at least one member of which is within household $h$, $w_{hjk}$ is considered equal to the weight of $j$ if $j$ was selected in the panel’s initial sample and $w_{hjk} = 0$ otherwise. The weight $w_h$ of $h$ can be written as

$$w_h = \frac{\beta_h}{\sum_{j=1}^{\beta_h} \sum_{k=1}^{m_j} \alpha_{hjk} w_{hjk}},$$

where $m_j$ is the number of individuals of household $j$ and

$$\sum_{j=1}^{\beta_h} \sum_{k=1}^{m_j} \alpha_{hjk} = 1.$$  

For the choice of parameters $\alpha_{hjk}$, it is possible to use $\alpha_{hjk} = \frac{1}{n_h}$ (Huang 1984; Ernst 1989), where $n_h$ is the number of members of the household $h$ who were in the original sample. Thus,

$$w_h = \frac{1}{n_h} \sum_{j=1}^{\beta_h} \sum_{k=1}^{m_j} w_{hjk}.$$  

### 4.4.1.3 Application

As an application of these two methods, changes in the composition of three households A, B and C of a panel between Wave 1 and Wave 4 are considered, and are illustrated in Figure 4.1 below. For example, it may be seen that in Wave 4, members of Household A are found in three different households: A2 joined Household C, A3 formed a new individual Household D and Household A welcomed a
new member E. It can also be seen that Member B3 of Household B has disappeared between the two waves.

**Figure 4.1** – Example of changes in the composition of households.

### 4.4.1.3.1 Calculation of weight with an equal household weighting scheme

The results of the calculations of household weights for Wave 4 are shown in Figure 4.2.a below. Let us explain, for example, the calculation of the weight of Household C. This household contains members from three households of Wave 1. Thus, its weight is equal to the average weight of these households: \((57 + 104 + 63)/3 = 74.7\).

<table>
<thead>
<tr>
<th>wave1</th>
<th>wave4</th>
</tr>
</thead>
<tbody>
<tr>
<td>household</td>
<td>HH member</td>
</tr>
<tr>
<td>A</td>
<td>A1</td>
</tr>
<tr>
<td>A</td>
<td>A2</td>
</tr>
<tr>
<td>A</td>
<td>A3</td>
</tr>
<tr>
<td>B</td>
<td>B1</td>
</tr>
<tr>
<td>B</td>
<td>B2</td>
</tr>
<tr>
<td>B</td>
<td>B3</td>
</tr>
<tr>
<td>B</td>
<td>B4</td>
</tr>
<tr>
<td>C</td>
<td>C1</td>
</tr>
<tr>
<td>C</td>
<td>C2</td>
</tr>
</tbody>
</table>

**Figure 4.2.a** – Example of weight calculation using the equal household weighting scheme.
4.4.1.3.2 Calculation of weight with the equal person weighting scheme

Figure 4.2.b below shows the results of calculations of weight using the equal person weighting approach. For example, Household B is linked to two households of Wave 1 (B and C), but three of its members were in this wave. Thus, the weight of B is equal to the average weight of these three members: \( \frac{104 + 104 + 63}{3} = 90.3 \).

<table>
<thead>
<tr>
<th>wave1 household</th>
<th>HH member</th>
<th>weight</th>
<th>wave2 household</th>
<th>HH member</th>
<th>( \beta_n )</th>
<th>( \eta_n )</th>
<th>( w_n = \frac{1}{\sum_{j=1}^{A_n} \sum_{h=1}^{A_n} w_{j,h}} )</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A1</td>
<td>57</td>
<td>A</td>
<td>A1</td>
<td>1</td>
<td>1</td>
<td>57</td>
<td>57.0</td>
</tr>
<tr>
<td>A</td>
<td>A2</td>
<td>57</td>
<td>A</td>
<td>E</td>
<td>1</td>
<td>1</td>
<td>57</td>
<td>57.0</td>
</tr>
<tr>
<td>A</td>
<td>A3</td>
<td>57</td>
<td>B</td>
<td>B1</td>
<td>2</td>
<td>3</td>
<td>271</td>
<td>90.3</td>
</tr>
<tr>
<td>B</td>
<td>B1</td>
<td>104</td>
<td>B</td>
<td>B2</td>
<td>2</td>
<td>3</td>
<td>271</td>
<td>90.3</td>
</tr>
<tr>
<td>B</td>
<td>B2</td>
<td>104</td>
<td>C</td>
<td>C1</td>
<td>2</td>
<td>3</td>
<td>271</td>
<td>90.3</td>
</tr>
<tr>
<td>B</td>
<td>B3</td>
<td>104</td>
<td>C</td>
<td>C1</td>
<td>3</td>
<td>3</td>
<td>224</td>
<td>74.7</td>
</tr>
<tr>
<td>B</td>
<td>B4</td>
<td>104</td>
<td>C</td>
<td>B4</td>
<td>3</td>
<td>3</td>
<td>224</td>
<td>74.7</td>
</tr>
<tr>
<td>C</td>
<td>C1</td>
<td>63</td>
<td>C</td>
<td>A2</td>
<td>3</td>
<td>3</td>
<td>224</td>
<td>74.7</td>
</tr>
<tr>
<td>C</td>
<td>C2</td>
<td>63</td>
<td>D</td>
<td>A3</td>
<td>1</td>
<td>1</td>
<td>57</td>
<td>57.0</td>
</tr>
</tbody>
</table>

N.B. The example of Household B shows that the two methods do not always yield the same result. According to Kalton and Brick (1994), when households of Wave 1 are selected with approximately the same probability, use of the equal household weighting scheme appears preferable.

4.4.2 The General Weight Share Method

The GWSM used in indirect sampling can also be used to estimate weights following changes of statistical units. Lavallée (2007) provides an example of the application of this method in Canada’s Survey of Labour and Income Dynamics (SLID). A simple application in agricultural survey panels is hereby proposed.

4.4.2.1 Definitions of links

Consider a household \( j \) of the first wave and a household \( h \) of the second wave. The link between these two families can be defined as follows: \( h \) is linked to \( j \) if at least one member of \( h \) was a member of \( j \).
4.4.2.2 Collecting data on the links

Data collection concerning the links can be performed in two steps:

1. for each household \( j \) of the first wave, all households \( h \) of the second wave linked to it must be identified;
2. for each linked household \( j \), data on the number of links with households of the first wave must be collected.

**Household links identified during Wave 1**

Information on these links can be collected from the following questions.

<table>
<thead>
<tr>
<th>Household Code: ...........</th>
<th>Full name of head of household: .........................</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many members have left your household between last year and this year? ..........</td>
<td></td>
</tr>
<tr>
<td>Among these members, how many have joined other households? ..........</td>
<td></td>
</tr>
<tr>
<td>Can you name these members and households they joined?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Member Code</th>
<th>The new Household Code</th>
<th>Name of the head of the new household</th>
</tr>
</thead>
<tbody>
<tr>
<td>............</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>............</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>............</td>
<td>..........................</td>
<td>..........................</td>
</tr>
<tr>
<td>............</td>
<td>..........................</td>
<td>..........................</td>
</tr>
</tbody>
</table>

*Table 4.3 – Questionnaire on the links of Wave 1 households.*

**Potential links of Wave 2 households**

The households of Wave 2 consist of the sampled households of Year 1 and new linked households. For each household, information should be collected on the Wave 1 households to which it is linked and the total number of links with the households from Wave 1.

The number of links can be estimated at individual or household level. The questions to be posed under each approach are provided in Table 4.4 below.
### Table 4.4 – Questionnaire on the linked households’ potential links

<table>
<thead>
<tr>
<th>Level</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household level</td>
<td>Are there members of your household that belonged to other households last year? (Yes/No) ............</td>
</tr>
<tr>
<td></td>
<td>In how many households did these members live? ..................</td>
</tr>
<tr>
<td>Individual level</td>
<td>How many members of your household belonged to other households last year? ............</td>
</tr>
</tbody>
</table>

### 4.4.2.3 Application

For the purposes of calculating the weights with the GSWM, a simulation is performed with data based on the changes in the composition of the three households A, B and C between two waves of a panel. An illustration of these changes is provided in Figure 4.3 below. In this simulation, a definition of the total number of links at the individual level is considered.

![Figure 4.3 – Example of changes in the composition of households, for a GWSM simulation.](image)

The household weights of Wave 2 are available in Figure 4.4 below. The method for calculating the weight is explained in detail in Chapters 2 and 3.
For instance, in Wave 2, Household C is linked to three households of Wave 1 (A, B and C). Thus, its initial weight is equal to the sum of the weights of these households (57+104+63=224). The final weight is the initial weight divided by the potential links: 224/3=74.6.

<table>
<thead>
<tr>
<th>household</th>
<th>HH member</th>
<th>weight</th>
<th>household</th>
<th>HH member</th>
<th>HH pot. Links</th>
<th>Initial weight</th>
<th>Final weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A1</td>
<td>57</td>
<td>A</td>
<td>A1</td>
<td>2</td>
<td>57</td>
<td>29</td>
</tr>
<tr>
<td>A</td>
<td>A2</td>
<td>57</td>
<td>B</td>
<td>B1</td>
<td>4</td>
<td>271</td>
<td>68</td>
</tr>
<tr>
<td>A</td>
<td>A3</td>
<td>57</td>
<td>B</td>
<td>B2</td>
<td>4</td>
<td>271</td>
<td>68</td>
</tr>
<tr>
<td>B</td>
<td>B1</td>
<td>104</td>
<td>B</td>
<td>B2</td>
<td>4</td>
<td>271</td>
<td>68</td>
</tr>
<tr>
<td>B</td>
<td>B2</td>
<td>104</td>
<td>C</td>
<td>C1</td>
<td>3</td>
<td>224</td>
<td>75</td>
</tr>
<tr>
<td>B</td>
<td>B3</td>
<td>104</td>
<td>C</td>
<td>B4</td>
<td>3</td>
<td>224</td>
<td>75</td>
</tr>
<tr>
<td>B</td>
<td>B4</td>
<td>104</td>
<td>C</td>
<td>A2</td>
<td>3</td>
<td>224</td>
<td>75</td>
</tr>
<tr>
<td>C</td>
<td>C1</td>
<td>63</td>
<td>D</td>
<td>A3</td>
<td>3</td>
<td>57</td>
<td>19</td>
</tr>
<tr>
<td>C</td>
<td>C2</td>
<td>63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.4 – GWSM simulation data (households).**

### 4.5 Changes in farms and weight calculation

In some countries, the use of large farms as statistical units is a common practice. Section 4.2 above analysed the changes that can take place in sampled farms from one year to another: area change (increase or decrease), disappearance of farms, changes in the population or in the economy that can affect the representativeness of a sample of farms, etc. This Section also examined some possible solutions that can apply when the changes are not random.

When the sampled farms’ sub-components are a subject of interest, certain changes may require farms’ weights to be re-computed if reliable cross-sectional estimates are to be obtained. This problem may arise if movements between the sub-components of sampled farms are possible. For example, all or a portion of a given farm sampled may belong to another statistical unit from one year to another. In this case, the method used to compute the weight of the new statistical units, based on the weights of the panel statistical units selected in Wave 1, can be used. These methods were described in detail in Section 4.4 above for the calculation of households’ weights. They can be easily adapted to the case of farms in which the movements of household members are considered as movements of the farm’s sub-components. The adaptation of these methods will be developed in this Section, and simulations will be presented.
4.5.1 Equal farm weighting scheme

Consider a farm $f$ with $m_f$ plots. For every farm $j$ that existed during the selection of the initial sample (Wave 1) and with at least one plot in the farm $f$, $w_{fj}$ is considered to be equal to the weight of $j$ if $j$ was selected in the panel’s initial sample and $w_{fj} = 0$ otherwise. The weight $w_f$ of farm $f$ can be written as:

$$w_f = \sum_{j=1}^{m_f} \alpha_{fj} w_{fj},$$

with

$$\sum_{j=1}^{m_f} \alpha_{fj} = 1,$$

in which $\beta_f$ is the number of farms in the initial sample that have a plot in the farm $f$.

As seen in Section 4.4 above, $\alpha_{fj} = \frac{1}{\beta_f}$ can be used. This gives

$$w_f = \frac{1}{\beta_f} \sum_{j=1}^{\beta_f} w_{fj}.$$

4.5.2 Equal plot weighting scheme

In this case, for any plot $p$ of a farm $j$ that existed during the selection of the initial sample (Year 1) and has at least one plot in the farm $f$, we consider $w_{fjp}$ to be equal to the weight of $j$ if $j$ was selected in the panel’s initial sample and $w_{fjp} = 0$ otherwise.
The weight $w_f$ of the farm $f$ can be written as:

$$w_f = \sum_{j=1}^{\beta_f} \sum_{p=1}^{m_j} \alpha_{fjp} w_{fjp},$$

where $m_j$ is the number of plots of farm $j$ and

$$\sum_{j=1}^{\beta_f} \sum_{p=1}^{m_j} \alpha_{fjp} = 1.$$

As seen above, $\alpha_{fjp} = 1/n_f$ can be used, where $n_f$ is the number of plots of the farm $f$ that were within the original sample. Thus,

$$w_f = \frac{1}{n_f} \sum_{j=1}^{\beta_f} \sum_{p=1}^{m_j} w_{fjp}.$$

### 4.5.3 The General Weight Sample Method

This subsection sets out a simple application of the GWSM method to the estimation of farms’ weights in panel surveys.

#### 4.5.3.1. Definitions of links

Consider a farm $j$ of Wave 1 and a farm $f$ of Wave 2. The link between these two farms may be described as follows: $f$ is linked to $j$ if at least one plot of $f$ was a plot of $j$. It is recalled that the term “plot” refers to a portion of a farm’s area.

In quantitative terms, the number of links between a farm $j$ of Wave 1 and a farm $f$ of Wave 2 may be estimated in two ways:

i. the number of plots of the farm $f$ that were included in the farm $j$ (number approach)

ii. the total area of plots of the farm $f$ that were included in the farm $j$ (size approach).
4.5.3.2 Collecting data on the links

Collecting data on the links can occur in two steps:

1. for each farm \( j \) of Wave 1, all farms \( h \) of Wave 2 linked to it must be identified;
2. for each linked farm \( h \), data on the number of links with the farms of Wave 1 must be collected.

**Links of the farms identified during Wave 1**

Information on these links can be collected through the questionnaire set out in Table 4.5 below.

<table>
<thead>
<tr>
<th>Farm code: ........</th>
<th>Full name of respondent: .........................</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many plots have left your farm between last year and this year?........</td>
<td>Among these plots, how many have joined other farms? ..........</td>
</tr>
<tr>
<td>Please give us more information on the new farms that they joined.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Plot code</th>
<th>The new farm code</th>
<th>Name of the head of the new farm</th>
</tr>
</thead>
<tbody>
<tr>
<td>..........</td>
<td>..................</td>
<td>...................................</td>
</tr>
<tr>
<td>..........</td>
<td>..................</td>
<td>...................................</td>
</tr>
<tr>
<td>..........</td>
<td>..................</td>
<td>...................................</td>
</tr>
<tr>
<td>..........</td>
<td>..................</td>
<td>...................................</td>
</tr>
</tbody>
</table>

*Table 4.5 – Questionnaire on the links of farms identified during Wave 1.*

**Potential links of Wave 2 farms**

The farms of Wave 2 consist of the farms sampled in the first year and of newly linked farms. For each farm of Wave 2, information should be collected on the farms in Wave 1 to which it is linked and on the total number of links with farms from Wave 1.
The questions that could be asked to estimate potential links under each approach (number or size) to estimate the number of links are provided in Table 4.6 below. Under the number approach, the number of links can be estimated at plot or at farm level.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Level</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number approach</td>
<td>Farm level</td>
<td>Are there plots of your farm that belonged to other farms last year? (Yes/no) ..........</td>
</tr>
<tr>
<td></td>
<td></td>
<td>To how many farms did these plots belong? ..................................</td>
</tr>
<tr>
<td></td>
<td>Plot level</td>
<td>Are there plots of your farm that belonged to other farms last year? (Yes/no) ..........</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How many plots of your farm belonged to other farms last year? ..........</td>
</tr>
<tr>
<td>Size approach</td>
<td>-</td>
<td>Are there plots of your farm that belonged to other farms last year? (Yes/no) ..........</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is the total size of these plots?</td>
</tr>
</tbody>
</table>

Table 4.6 – Questionnaire on the linked farms’ potential links.

4.5.4. Application

For the purposes of these three methods, the example of the movements of three farms’ plots between two waves of a panel will be used, as shown below. Generally, in practice, such a great number of plot movements between two waves are not common: this example was constructed as a simulation.

It can be seen that Plot B3 of Farm B disappeared between the two waves: this phenomenon may be related, for example, to degradation of the plot or fallow. In Wave 2, it is assumed that Plots E, F, G and H were not part of the sample during Wave 1. Their presence can be explained by an expansion of the agricultural holdings through the purchase or lease of new plots.
For each method, the weight estimation of Wave 2 farms are given below.

### 4.5.4.1 Application of the equal farm weighting scheme

**Example**: Farm B of Wave 2 contains plots of two farms of Wave 1 (B and C). Thus, its weight is equal to the mean of the weights of these farms: \((34+17)/2 = 25.5\).

<table>
<thead>
<tr>
<th>wave1</th>
<th>wave2</th>
</tr>
</thead>
<tbody>
<tr>
<td>farm</td>
<td>plot</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>A</td>
<td>A1</td>
</tr>
<tr>
<td>A</td>
<td>A2</td>
</tr>
<tr>
<td>A</td>
<td>A3</td>
</tr>
<tr>
<td>B</td>
<td>B1</td>
</tr>
<tr>
<td>B</td>
<td>B2</td>
</tr>
<tr>
<td>B</td>
<td>B3</td>
</tr>
<tr>
<td>B</td>
<td>B4</td>
</tr>
<tr>
<td>C</td>
<td>C1</td>
</tr>
<tr>
<td>C</td>
<td>C2</td>
</tr>
</tbody>
</table>

**Figure 4.5.a – Example of weight calculation using the equal farm weighting scheme.**
4.5.4.2 Application of the equal plot weighting application

Example: Farm C of Wave 2 contains three plots that were in Wave 1 (C1, B4, A2). Thus, its weight is equal to the mean of the weights of these plots: (17 + 34 + 12) / 3 = 21.

<table>
<thead>
<tr>
<th>wave1</th>
<th>farm</th>
<th>plot</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A1</td>
<td>12</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>A2</td>
<td>12</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>A3</td>
<td>12</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>B1</td>
<td>34</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>B2</td>
<td>34</td>
<td>C</td>
</tr>
<tr>
<td>B</td>
<td>B3</td>
<td>34</td>
<td>C</td>
</tr>
<tr>
<td>B</td>
<td>B4</td>
<td>34</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>C1</td>
<td>17</td>
<td>D</td>
</tr>
<tr>
<td>C</td>
<td>C2</td>
<td>17</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.5.b – Example of weight calculation using the equal plot weighting scheme.

4.5.4.3. Application of the General Weight Sample Method

For application of the GWSM, simulation results will be shown below for each approach to the estimation of the number of links.

The number approach

The number of links at plot level will now be considered.

Example: Farm B of Wave 2 has three links with two farms of Wave 1 (B and C). Thus, its initial weight is: 34 + 34 + 17 = 85. It is also linked to the farm that contained Plot F but that was not sampled in Wave 1; therefore, Farm B has a total of four links and its final weight is equal to 85 / 4 = 21.3.
The size approach

Example: The initial weight of Farm B is: 34 x 7.4 + 34 x 15.6 + 17 x 19.2 = 1,108. B is also linked to the farm that contained Plot F but that was not sampled in Wave 1. Considering that the size of Plot F is 5.1 acres, the total weighted links (i.e. weighted according to the plot sizes) of Farm B is 7.4 + 15.6 + 19.2 + 5.1 = 47.3, and its final weight is equal to 1108 / 47.3 = 23.4.
Optimal sampling

5.1 Introduction

The sampling strategy for parameter estimation in survey sampling is defined by the couple sampling design and estimator. If a randomization approach to the inference is adopted, the design must envisage a random selection, so that for each sub-population $s$ of the reference population $U$, a selection probability $p(s) \geq 0$ with $\sum_{s \in U} p(s) = 1$ is given. Standard selection schemes are one-stage or multi-stage sampling designs, and simple (stratified or non-stratified) random sampling designs. When the probabilities are proportional to a given size measure, PPS sampling designs are implemented. Given the selection scheme, a sampling design (i.e. the selection probability distribution of the subpopulations of $U$) can be defined when the inclusion probabilities of the units are fixed. For example, the Stratified Simple Random Sampling without replacement (SSRS) design is defined by the probabilities given by the ratio of the sample size and the population size at the stratum level, and by selecting a fixed number of units in each stratum (selection scheme).

The definition of the inclusion probabilities is a fundamental step in defining an efficient sampling strategy: that is, a strategy that, given a sample size, produces the lowest sampling errors possible.

This Chapter focuses on the definition of the inclusion probabilities in the case of SSRS without replacement. This, and two-stage sampling designs, are commonly used in the official statistics produced by National Statistical Offices. Other types of
Fixing the inclusion probabilities of SSRS means defining the stratum’s sample sizes; thus, the problem is denoted as a sample allocation (in the strata) or as an allocation problem. The literature on sampling has devoted much attention to sample allocation. When one target parameter is to be estimated for the overall population, the optimal allocation in stratified samplings can be performed (Cochran, 1977). In particular, the Neyman sample allocation minimizes the variance of the subject depending on a given budget (the cost-constraint optimization problem). Alternatively, reversing the problem, an allocation that minimizes costs can be performed, subject to a given sampling error threshold (the precision constraint optimization problem).

When more than one parameter and domain estimation are to be estimated, the allocation is no longer optimal. If the strata are domains of interest, the Neyman allocation may cause certain strata to undergo great variation, because the strata are not domains of estimation. The Neyman allocation is no longer optimal, and a more suitable rough sample allocation may be the equal sample allocation. Equal allocation is efficient when estimating stratum (or specific domain) parameters, but may lead to a much greater variance of the estimator than that obtained through the Neyman allocation for the overall population. In several practical applications, a compromise allocation must be performed.

In multivariate cases, where more than one characteristic of each sampled unit must be measured, the optimal allocation for individual characteristics are of little practical use, unless the characteristics under study are highly correlated. This is because an allocation that is optimal for one characteristic is generally far from optimal for others. The multidimensionality of the problem also leads to a compromise allocation method being defined (Khan et al. 2010), with a loss of precision compared to the individual optimal allocations. Several authors have discussed various criteria for obtaining a plausible compromise allocation: see, for example, Kokan and Khan (1967), Chromy (1987), Bethel (1989) and Choudhry et al. (2012). In Section 5.2 below, the compromise allocation is formalized as a particular optimization problem in which the concept of optimality differs from the cost-constraint or precision-constraint univariate and the uni-domain optimization problem. The compromise allocation takes into account only the parameters of interest of the directly sampled population. Section 5.3 below extends the sample allocation procedure, also considering the parameters of interest in the indirect sampled population. Section 5.4 outlines a special context of sampling allocation, taking into account a two-phase sampling design to observe the indirect sample. Section 5.5 is devoted to numerical examples on real data.
5.2 Sample allocation in the stratified simple random sampling design

The compromise allocation is expressed as an optimization problem. Let $J$ be both the overall target population and the size of this population in which $J$ is either $H$ (household) or $F$ (farm), and let $J_r$ be both the population and the size of the $r$th stratum ($r=1,\ldots,R$). The parameter of interest is the total

$$Y_{J,v} = \sum_j Y_{Jr,v},$$

where

$$Y_{Jr,v} = \sum_{j \in J_r} y_{j,v}$$

is the stratum total computed summing up the values $y_{j,v}$ of each unit $j$ of the variable $y_v$ ($v=1,\ldots,V$).

We consider the unbiased Horvitz-Thompson (H-T) estimator

$$\hat{Y}_{J,v} = \sum_{j \in s_J} y_{j,v} w_j,$$

in which $w_j = 1/\pi_j$.

Let $s_{Jr}$ indicate the random sample in the stratum $r$ and let $J_r$ and $n_{Jr}$ denote, respectively, the number of units in the stratum $r$ in the population and the number of units in the sample. For instance, if $J=H$, then $J_r = H_r$ denotes the number of households in the stratum $r$, and $n_{Jr} = n_{Hr}$ indicates the size of the stratum’s sample. Similarly, if $J=F$, then $J_r = F_r$ and $n_{Jr} = n_{F_r}$ denotes the number of farms in the population and in the sample of stratum $r$.

In the SSRS design, $w_j = \frac{J_r}{n_{Jr}}$ for $j \in s_{Jr}$. Therefore, the H-T estimator can be reformulated as

$$\hat{Y}_{J,v} = \sum_{r=1}^R \sum_{j \in s_{Jr}} \frac{J_r}{n_{Jr}} y_{j,v},$$
Example 1: The Use of Symbology

Due to the additional difficulty discussed in this Chapter, the symbology used above is more complex than that described in the previous chapters. To illustrate it, let us consider a survey on households. In this case, \( J = H \) and \( j = h \). Suppose, further, that the survey has two target objectives with \( V = 2 \).

- **First**: the total of employed people. In this case, \( v = 1 \), the symbol \( Y_{J,v} \) is equal to \( Y_{H,1} \). The elementary variable \( y_{j,v} = y_{h,1} \) indicates the number of employed persons in the household \( h \) and the total \( Y_{Jr,v} = Y_{Hr,1} \) denotes the number of employed persons in the stratum \( r \).

- **Second**: the total number of employed persons in agriculture. In this case, \( v = 2 \), the symbol \( Y_{J,v} \) is equal to \( Y_{H,2} \), the elementary variable \( y_{j,v} = y_{h,2} \) indicates the number of individuals employed in agriculture in the household \( h \) and the total \( Y_{Jr,v} = Y_{Hr,2} \) denotes the number of individuals employed in agriculture in the stratum \( r \).

Large-scale surveys usually produce estimates for subpopulations or domains \( J_d \) \((d=1, \ldots, D)\). The domain estimation is easily addressed, with an H-T estimator using the domain-specific study variable \( y_q \), where \( q \) is a couple of \((v, d)\), and the value \( y_{j,q} \) is defined as

\[
y_{j,q} = \begin{cases} y_{j,v}, & \text{if unit } j \in J_d \\ 0, & \text{otherwise} \end{cases}
\]

Then, the domain estimates are treated as estimates of \( \hat{Y}_{J,q} = \sum_{r} \sum_{j \in J_r} (J_r/n_r) y_{j,q} \) at the overall level. Below, the estimate of \( Y_{J,v} \), without loss of generality will be considered.
The variance of \( \hat{Y}_{j,v} \) is given by

\[
V(\hat{Y}_{j,v}) = \sum_{r=1}^{R} \frac{V_{j_r,v}}{n_{j_r}} + V_{0j,v}, \tag{5.2.1}
\]

in which

- \( V_{j_r,v} = J_r^2 \sigma_{j_r,v}^2 \),

being

\[
\sigma_{j_r,v}^2 = \frac{\sum_{j \in r} [y_{j,v} - (1/J_r \sum_{j \in r} y_{j,v})]^2}{(J_r - 1)} \tag{5.2.2a}
\]

the variance of the variable \( y_v \) in the stratum \( r \) and

- \( V_{0j,v} = -\sum_{r=1}^{R} J_r \sigma_{j_r,v}^2 \) \tag{5.2.2b}

is a fixed part that does not depend on stratum sample sizes.

Finally, the coefficient of variation is defined as

\[
CV(\hat{Y}_{j,v}) = \sqrt{V(\hat{Y}_{j,v})/\hat{Y}_{j,v}}.
\]

In the allocation problem, the terms \( n_{j_r} \) are unknown and must be planned. Chromy (1987), Bethel (1989) and Choudhry et al. (2012) give a mathematical formalization to the compromise allocation, according to the optimization problem:

\[
\begin{aligned}
\min & \sum c_{j_r} n_{j_r} \\
\sum_{r} V_{j_r,v}/n_{j_r} + V_{0j,v} & \leq V_{j,v}^* \quad \forall v = 1, \ldots, V, \\
1 \leq n_{j_r} & \leq J_r
\end{aligned} \tag{5.2.3}
\]

where \( c_{j_r} \) denotes the cost per unit of collecting the information in the stratum \( r \) and \( V_{j,v}^* \) is the variance threshold that denotes the maximum value acceptable for the
variance $V(\hat{Y}_{j,v})$ of the estimate $\hat{Y}_{j,v}$. For example, if $\bar{V}_{j,v} = 0.10$, the minimum sample size must guarantee that the total $\hat{Y}_{j,v}$‘s sample variance is lower than or equal to 0.10.

**BOX 5.2**

**EXAMPLE 2: THE DEFINITION OF THE OPTIMAL ALLOCATION**

Consider the problem described in Example 1 above and suppose that the variance threshold $V_{j,v}^*$ for the first and second parameters of interest are

- $V_{H,1}^* = 0.10$ and 
- $V_{H,2}^* = 0.15$.

Problem (5.2.3) finds the minimum cost solution, guaranteeing that

- the sample variance of the estimate of the number of individuals employed $V(Y_{H,1})$ is lower than or equal to 0.10, and at the same time that

the sample variance of the estimate of the number of individuals employed in agriculture $V(Y_{H,2})$ is lower than or equal to 0.15.

Note that if $c_{Jr} = c$ and $V=1$, the optimization problem (5.2.3) coincides with the standard precision-constraint optimization problem.

Furthermore, when $\sigma_{Jr,v}^2 = \sigma_J^2$ and $V=1$, the optimal solution coincides with the proportional-to-stratum-population-size allocation.

Several algorithms proposed in literature solve Problem (5.2.3). Among these, Chromy (1987) developed an algorithm that is suitable for automated spreadsheets. The algorithm is implemented by the MAUSS-R software (which can be downloaded at http://www.istat.it/it/strumenti/metodi-e-strumenti-it/strumenti-di-progettazione/mauss-r). The software requires the installation of “R”, a free software environment for statistical computing (download available at http://www.cran.r-project.org). More recently, Choudhry et al. (2012) noted that the objective function is a convex separable function. They deal with the problem as Non-Linear Programming, and use the SAS proc NLP with the N-R option.
In practical applications, the input parameters of the optimization problem are \( \sigma_{J_r,v}^2 \), \( Y_{J,r} \) and the variance thresholds are generally given as \( V_{J,r}^* = \left[ CV^*(\hat{Y}_{J,v}) \right]^2 Y_{J,r}^2 \), being \( CV^*(.) \) the coefficient of variation’s thresholds. The algorithm’s outputs are the stratum sample sizes and the expected CVs of the estimates. Usually, expected CVs are much lower than the respective \( CV^* \)'s, except in certain cases where these values are close.

Finally, the parameters \( \sigma_{J_r,v}^2 \) and \( Y_{J,r} \) are unknown, because they are the survey’s targets. These values must then be replaced with some estimates, and the estimates must be treated as true values. A common strategy is to use data from previous surveys when the variable \( y_v \) was collected and to perform an estimation procedure. A second strategy is to use the totals’ variances of proxy variables known at the population level. Thus, in the estimation phase, the CVs observed should be different from those expected. The difference depends upon the accuracy of the input parameters.

### 5.2.1 Extension to the case of a stratified multi-stage sampling design

Section 5.2 introduced the sample allocation for the SSRS design. The general approach is easily adapted to more complex designs such as the Multistage Stratified PPS Sampling Design, in which \( n_{J_r} \) is the elementary unit sample size selected in each stratum. This Section outlines the main steps and details some technical issues that are essential to implementation of the sample allocation. Extensive descriptions of the sample allocation for general sampling designs are shown by FAO (2014).

**Sampling design**

The design is defined as follows:

- \( m_{1r} \) PSUs are selected in the stratum \( J_r \). The \( i-th \) PSU is selected without replacement, with inclusion probability \( \pi_{1i} = m_{1r} J_{ri} J_r \). \( J_{ri} \), being the number of elementary units (households or farms) in the \( i-th \) PSU;

- assume that the elementary units of the selected PSU are split into \( M_{2r} \) Enumeration Areas (EAs), each of which has approximately the same number
$\bar{J}_r$ of elementary units. This number represents the workload (in terms of interviews) that an enumerator can perform in a given time interval (e.g. one week). Therefore,

$$M_{2r} \approx \frac{J_{ri}}{\bar{J}_r};$$

- $m_{2r}$ EAs are selected in each sampled PSU, without replacement and with equal probability. All the elementary units of each selected EA are observed. Thus, $\bar{n}_{Jr} = m_{2r} \bar{J}_r$ elementary units are observed in each PSU. The variable $m_{2r}$ is determined such that

$$m_{1r} m_{2r} \bar{J}_r = n_{Jr},$$

thus,

$$m_{2r} = \frac{n_{Jr}}{m_{1r} \bar{J}_r}.$$  

This sampling strategy ensures that each elementary unit in the stratum $r$ has an equal probability of being included in the sample:

$$\pi_j = \pi_{1i} \frac{m_{2r}}{M_{2ri}} = \frac{n_{Jr}}{\bar{J}_r}.$$  

Indeed, by means of a simple substitution of $\pi_{1i}$, $m_{2r}$ and $M_{2ri}$ respectively with

$$m_{1r} \frac{J_{ri}}{\bar{J}_r}, \frac{n_{Jr}}{m_{1r} \bar{J}_r} \text{ and } \frac{J_{ri}}{\bar{J}_r},$$

$$\pi_j = m_{1r} \frac{J_{ri}}{\bar{J}_r} \times \frac{n_{Jr}}{m_{1r} \bar{J}_r} \times \frac{\bar{J}_r}{J_{ri}} = \frac{n_{Jr}}{J_{r}}.$$
**Sampling variance and the optimization problem**

According to Kish (1966), the sampling variance of the above strategy may be determined according to:

\[
V(\hat{Y}_{J,v}) = \sum_{r=1}^{R} \frac{V_{Jr,v}}{n_{Jr}} + V_{0J,v},
\]

where

\[
V_{Jr,v} = J_r^2 \bar{\sigma}_{Jr,v}^2,
\]

and

\[
V_{0J,v} = -\sum_{r=1}^{R} J_r \bar{\sigma}_{Jr,v}^2
\]

in which

\[
\bar{\sigma}_{Jr,v}^2 = \sigma_{Jr,v}^2 [1 + \rho_r (m_{2r} \bar{J}_r - 1)],
\]

\(\rho_r\) being the coefficient of intra-class correlation in the stratum.

Therefore, having an estimation of \(\rho_r\) and having defined the quantities

- \(m_{2r}\) and
- \(\bar{J}_r\),

on the basis of costs and logistics considerations, then

- the stratum sample sizes in terms of elementary units \(n_{Jr}\) are defined by solving the problem (5.2.3), in which the modified variances \(\bar{\sigma}_{Jr,v}^2\) substitute the variances \(\sigma_{Jr,v}^2\)
- the number of PSUs per stratum \(m_{1r}\) is defined by

\[
m_{1r} = \frac{n_{Jr}}{m_{2r} \bar{J}_r}.
\]
The PSUs are selected in each stratum by means of a PPS design without replacement.

**Remark 5.2.1:** The allocation problem can be repeated for different values of $m_{2r}$.

### 5.3 Sample allocation in the SSRS design, considering the indirect sample selection

The sample allocation of the directly sampled population can take into account the second step of the indirect sample selection. To do so, auxiliary information on the indirect population must be available at the design stage. The general case involving the populations $H$ and $F$ for different qualitative levels of auxiliary information will be considered. It will be shown that the sampling approach proposed herein meets the requirements of managing the sample allocation of the two linked populations at the same time. This is a first step in defining a global sampling strategy for the socioeconomic and environmental parameters of different populations. Details on the cost function for a complete definition of the optimization problem are given in FAO (2014).

There are two main cases:

**Case A:** There is a Master Sampling Frame (MSF) in which the direct links between households and farms are available. In this case, the frame enables identification of the links between the populations observed in Steps 1 and 2.

**Case B:** The sampling frame that refers to the population observed in Step 1 has some auxiliary statistics referred to the population observed in Step 2 (the indirectly observed population).
5.3.1 Case A: Master Sampling Frame

Direct links between the directly and indirectly sampled units are available in the sampling frame. In this case, the conditions are ideal for planning a sample allocation according to an integrated approach.

A.1 - From households to farms

The problem may be defined as follows:

1. There are $V$ totals of interest:

$$Y_{H,1},...,Y_{H,v},...,Y_{H,V_H},Y_{F,V_H+1},...,Y_{F,V}.$$  

The first $V_H$ totals refer to the population of households, while the second $V_H - V$ totals refer to the population of farms.

2. There are $V$ variance thresholds:

$$V^*_H,*,V^*_{H,v},*,V^*_{H,V_H},*,V^*_{H,V_H+1},*,V^*_{F,V},*.$$  

that define the sample variances' upper bound:

$$V(\hat{Y}_{H,1}),...,V(\hat{Y}_{H,v}),...,V(\hat{Y}_{H,V_H}),V(\hat{Y}_{F,V_H+1}),...,V(\hat{Y}_{F,V}).$$

3. The sample strategy is based on a two-step scheme.

   a. First, the households are sampled, to collect the variables $Y_{h,v}$ and to build the estimates $\hat{Y}_{H,v}$ ($v=1,...,V_H$).

   b. Second, the farms are sampled through the indirect sampling mechanism, to collect the variables $Y_{f,v}$ and compute the estimates $\hat{Y}_{F,v}$ ($v=V_H+1,...,V$).
In this context, the optimization problem (5.2.3), which ascertains the solution entailing the minimum cost, can be reformulated as:

\[
\begin{align*}
\min & \sum_r c_{H_r} n_{H_r} \\
\text{s.t.} & \sum_r V_{H_r,v} n_{H_r} + V_{0H,v} \leq V_{H,v}^* & \forall v = 1, \ldots, V_H \\
& \sum_r V_{F_r,v} n_{H_r} + V_{0F,v} \leq V_{F,v}^* & \forall v = V_H + 1, \ldots, V \\
& 1 \leq n_{H_r} \leq H_r
\end{align*}
\]

(5.3.1)

in which

- \( V_{H_r,v} = H_r^2 \sigma^2_{Hr,v} \), \( V_{0H,v} = -\sum_{r=1}^R H_r \sigma^2_{Hr,v} \), with
  \[
  \sigma^2_{Hr,v} = \frac{1}{H_r-1} \sum_{h=1}^{H_r} \left[ y_{h,v} - \left( \frac{1}{H_r} \sum_{h=1}^{H_r} y_{h,v} \right) \right]^2;
  \]

- \( V_{F_r,v} = H_r^2 \sigma^2_{Zr,v} \), \( V_{0F,v} = -\sum_{r=1}^R H_r \sigma^2_{Zr,v} \),
  \[
  \sigma^2_{Zr,v} = \frac{1}{H_r-1} \sum_{h=1}^{H_r} \left[ z_{h,v} - \left( \frac{1}{H_r} \sum_{h=1}^{H_r} z_{h,v} \right) \right]^2,
  \]

where \( z_{h,v} = \sum_f (l_{h,f}/L_f) y_{f,v} \), with \( L_f = \sum_h l_{h,f} \).

Given \( \hat{Z}_{H,v} = \sum_f \sum_{H_r} (H_r/n_{H_r}) z_{h,v} \), we have \( \hat{Y}_{F,v} = \hat{Z}_{H,v} \), and the optimization problem replaces the variance \( \hat{Y}_{F,v} \) with the variance \( \hat{Z}_{H,v} \).
Consider the problem described in Case A.1. The method applied for building the \( z_v \) variable referred to households to estimate the farm parameter \( Y_{F,v} \) depends on the available links.

- **Household – Farm by individual link:**
  \[
  \frac{l_{h,f}}{L_f} = \begin{cases} 
  \frac{M_{h,f}}{M_f}, & \text{when } h \text{ is linked to } f \\
  0, & \text{otherwise}
  \end{cases}
  \]
  where \( M_{h,f} \) is the number of workers of the farm \( f \) belonging to the household \( h \);
  \( M_f \) the number of workers in farm \( f \) (total number of links).

- **Household – Farm by household link:**
  \[
  \frac{l_{h,f}}{L_f} = \begin{cases} 
  \frac{1}{H_f}, & \text{when } h \text{ is linked to } f \\
  0, & \text{otherwise}
  \end{cases}
  \]
  where \( H_f \) is the number of households with one or more individuals working in the farm \( f \) (total number of links).

### A.2 - From farms to households

The problem can be defined in the same way as in Case A.1, where \( F \) and \( H \) exchange roles so that the sample allocation can solve the following problem:

\[
\begin{align*}
\min \sum_r c_{Fr} n_{Fr} \\
\sum_r V_{Fr,v}/n_{Fr} + V_{0F,v} & \leq V_{F,v}^* & \forall v = 1, \ldots, V_F \\
\sum_r V_{Hr,v}/n_{Fr} + V_{0H,v} & \leq V_{H,v}^* & \forall v = V_F + 1, \ldots, V \\
1 & \leq n_{Fr} \leq F_r
\end{align*}
\]
in which

- $V_{Fr,v} = F_r^2 \sigma_{Fr,v}^2$, $V_{0F,v} = -\sum_{r=1}^R F_r \sigma_{Fr,v}^2$, being

  $\sigma_{Fr,v}^2 = \frac{1}{F_r - 1} \sum_{h=1}^{F_r} \left[ y_{f,v} - \left( \frac{1}{F_r} \sum_{f=1}^{F_r} y_{f,v} \right) \right]^2$.

- $V_{Hr,v} = F_r^2 \sigma_{Zr,v}^2$, $V_{0H,v} = -\sum_{r=1}^R F_r \sigma_{Zr,v}^2$,

  $\sigma_{Zr,v}^2 = \frac{1}{F_r - 1} \sum_{f=1}^{F_r} \left[ z_{h,v} - \left( \frac{1}{F_r} \sum_{f=1}^{F_r} z_{h,v} \right) \right]^2$.

For the sake of brevity, focus will be placed upon the household parameters $Y_{H,1}, ... , Y_{H,v}, ... , Y_{H,V_H}$.

In the optimization problem, each $\hat{Y}_{H,v} = \sum_{r} \sum_{s_{Fr}} (H_r/n_{Fr}) y_{h,v}$ is replaced with $\hat{Z}_{F,v} = \sum_{r} \sum_{s_{Fr}} (F_r/n_{Fr}) z_{f,v}$, where $z_{f,v} = \sum_{h} (l_{f,h}/L_h) y_{h,v}$, and $\hat{Y}_{H,v} = \hat{Z}_{F,v}$. In the optimization problem, the variance of $\hat{Y}_{H,v}$ is replaced by the variance of $\hat{Z}_{F,v}$. In Section 5.4. below, $V_{Fr,v}^Z$ is used as an alternative notation of $V_{Hr,v}$, and $V_{0F,v}^Z$ to indicate $V_{0H,v}$.
BOX 5.4
EXAMPLE 4: FROM FARMS TO HOUSEHOLDS: BUILDING THE Z VARIABLES FOR THE SAMPLE ALLOCATION

Consider the problem described in Case A.2. The method to be applied in building the $z_v$ variable referred to farms to estimate the household parameter $Y_{H,v}$ depends on the links available.

- Farm – Household by individual link:

$$\frac{I_{f,h}}{L_h} = \begin{cases} 
M_{h,f}/M_h & \text{when } f \text{ is linked to } h \\
0 & \text{otherwise}
\end{cases}$$

where $M_{h,f}$ is the number of workers of the farm $f$ belonging to the household $h$, and $M_h$ is the number of workers engaged in the agricultural activity of household $h$ (total number of links).

- Farm – Household by household link:

$$\frac{I_{f,h}}{L_h} = \begin{cases} 
1/F_h & \text{when } u_f \text{ is linked with } u_h \\
0 & \text{otherwise}
\end{cases}$$

where $F_h$ is the number of farms with at least one individual of household $h$ (total number of links).

**Remark 5.3.1.** The links shown in Examples 3 and 4 above are standard Boolean links. Deville and Lavallée (2006) and Lavallée and Labelle-Blanchet (2013) propose the use of optimal weighted links that reduce the variance of an indirectly investigated variable. The linkage weights are optimal for one parameter of estimation, and using can be attractive when tackling a univariate indirect sample allocation.
5.3.2 Case B: Sampling frame with auxiliary statistics referred to the indirectly observed population

In many real contexts, data integration is not possible, because the record linkage process does not provide good linkage or simply because a frame population does not exist. If the totals or estimated totals of the indirect population are available, these can be used to predict the $z$ values employed to implement the allocation of the indirectly sampled population.

B.1 - From households to farms

Let $Y_{F,v}$ be a parameter of interest of the farm population. As usual, it is assumed that the parameters and the population variance are known (see Section 5.2), although in practice, an estimated value or a total of correlated variables are used. From the sample allocation, we must define $z_{h,v}$ such that $\hat{Z}_{H,v} = \sum_r \sum_{H_r} (H_r/n_{H_r}) z_{h,v}$ is equal to $\hat{Y}_{F,v}$. Without household-farm links, a prediction of the $z_{h,v}$ can be built. The validity of the prediction will depend on whether the following reasonable assumptions hold:

- First assumption: reliable farm values $Y_{F,v}^{d_h}$ on the domains $F_d$ are available, and which $h \in F_d$ is known from the household sampling frame. Typically, $F_d$ is a geographical domain and the households and farms belong to the same domain at least at country level.

- Second assumption: for each household $h$, the $M_h$ values are available, being the number of workers in the household $h$ or a proxy variable (e.g. the number of household components of working age) known in the household sampling frame.

- Third assumption: each worker belonging to $h \in F_d$ works in the domain $F_d$. 
Consider the assumptions of Case B.1. From a farm survey (or farm register) reliable values are available for total crop, livestock or other parameters correlated to the phenomena of interest in the ISF for farms. The value $Y_{F,v}^{\text{region}}$ is at the regional level. In the household sampling frame, the place where each household $h$ is located and the number of workers $M_h$ engaged in agricultural activity are known. For example,

- in Region A, there is $Y_{F,v}^{A} = 100,000$;
- Household $h$ living in Region A has $M_h = 5$ workers;
- the total number of workers in Region A is $M_{H}^{A} = \sum_{h \in F_{A}} M_h = 330,000$.

The prediction of $z_{h,v}$ is given by

$$z_{h,v} = \frac{5}{330,000} \times 100,000 = 1.51.$$ 

B.2 - From farms to households

Let $Y_{H,v}$ be a parameter of interest of the household population. As usual, it is assumed that the parameters and the population variance (see Section 5.2 above) are known, although in practice, an estimated value or a total of correlated variables are used. From the sample allocation, $z_{f,v}$ must be defined such that $\hat{Z}_{F,v} = \sum_r \sum_{s_{Fr}} (F_r/n_{Fr}) z_{f,v}$ is equal to $\hat{Y}_{H,v}$. Without household-farm links, a prediction of the $z_{f,v}$ can be constructed.
The validity of the prediction will depend on whether the following reasonable assumptions hold:

- First assumption: reliable household values $Y_{H,v}^d$ (referring to agricultural activities) on the $H_d$ domains are available, and it is known, from the farm sampling frame, which $f \in H_d$ for certain $d$. Typically, $H_d$ is a geographical domain and the households and farms belong to the same domain at least at country level.

- Second assumption: for each farm $f$, the $M_f$ value is available, being the number of workers or a proxy variable such as a farm size measurement known in the farm sampling frame.

- Third assumption: each worker working in the domain $H_d$ lives in the domain $H_d$. 
BOX 5.6
EXAMPLE 6: FROM FARMS TO HOUSEHOLDS: BUILDING THE PREDICTION OF THE Z VARIABLES FOR THE SAMPLE ALLOCATION

Consider the assumptions of Case B.2. From a household survey (or a household register) reliable values are available for the total rural households, the total workers (by gender) engaged in agricultural activity, the rural household consumer expenditure or other parameters correlated to the phenomena of interest in the ISF for households. The value $Y_{H,v}^{region}$ is at regional level. In the farm sampling frame, the location of each farm $f$ and the number of workers $M_f$ are known. For example,

- in Region A, $Y_{H,v}^{A} = 100,000$;
- farm $f$ in Region A has $M_f = 50$ workers;
- the total number of workers in Region A is $M_F^A = \sum_{s \in F_A} M_f = 330,000$.

The prediction of $z_{f,v}$ is given by

$$z_{f,v} = \frac{50}{330,000} \times 100,000 = 15.15.$$  

5.4 Special case: subsampling large farms’ households

The optimal sample allocation controls the direct sample size, while the indirect sample size can be highly diverse. This occurs especially when the observational process from $F$ to $H$ is considered and the size of the farms in terms of workers is highly variable. One solution to control the number $n_H$ of sampled households is to define farm strata that are homogeneous in terms of numbers of workers. Usually, small strata are obtained (i.e., $F_r$ small) with larger farms and large strata (i.e. $F_r$ large) with smaller farms, and there is a direct relationship between $n_F$ and $n_H$.

Fine-tuning the $CV^*$ enables achievement of the dimensions sought.
For highly skewed farm distributions, very large farms can be excluded from the design and can be surveyed by means of a different design.

Another approach to control $n_H$ is to subsample in the set of selected households according to a Two-Phase Sampling design (Särndal et al., 1992), as shown in Chapter 2.

In this case, the variance of $\hat{Z}_{F,v} = \sum_r \sum_{i \in F_r} (F_i/n_{F_r}) z_{f,v}$ that is useful for the sample allocation (see Section 5.2) is

$$V(\hat{Z}_{F,v}) = \sum_r (V_{F,r,v}/n_{F,r}) + V_{0,F,v} + V_{2,F,v},$$

where $V_{2,F,v}$ is the variance of the variable $z_v$ (related to the variable $y_v$), due to the subsampling of households. A practical approach to the optimal allocation is to use an estimation of $V_{2,F,v}$ according to the data of previous surveys, and to set up the variance threshold of the two-phase sample $V_{2H,v}^*$ of the optimization problem as

$$V_{2H,v}^* = V_{H,v}^* - V_{2,F,v},$$

where $V_{H,v}^*$ is the variance threshold referred to the estimate without subsampling.

5.5 Numerical examples

The following numerical examples focus on the Multidomain and Multivariate (MM) method implemented by Systems (5.3.1) or (5.3.2). Two applications of sample allocation in the integrated framework are shown. The first is based on real survey data concerning Districts 7, 8 and 9 of the Gaza Province, Mozambique. Two integrated populations are taken into account: one of households (more precisely, heads of household) and one of farms. The two populations are merged, thus defining an artificial integrated database. The aim of the numerical examples is to verify whether the sampling strategy is improved when the indirectly sampled population is taken into account in the sample allocation.

The second application uses the database of agricultural households from Burkina Faso’s General Census of Agriculture. In the numerical example, a context in which there is no integrated survey framework is simulated. The sample allocation exploits the household sampling framework along with reliable farm statistics. The aim of the
numerical example is to show the main steps to be followed in performing the allocation.

Both simulation studies use the MAUSS-R software (available for download at http://www.istat.it/it/strumenti/metodi-e-strumenti-it/strumenti-di-progettazione/mauss-r), implemented in R language (Java is required for the user-friendly interface). The software produces the sample allocation, implementing the optimization problem introduced in Sections 5.2 and 5.3 above. Once installed, the software features a comprehensive English-language user guide.

5.5.1 Establishing the IT procedure

The software for the optimal sample allocation requires definition of two specific input datasets. The first dataset includes information on the stratified sampling design. Each record identifies a stratum, and the following mandatory variables must be filled: the stratum identifier; the stratum’s population size $J_r$; the stratum’s mean values $\bar{y}_{J,vr}$ $(v = 1, \ldots, V)$ and, if the indirect sampling is taken into account, the stratum’s mean values $\bar{Z}_{J,vr}$; the stratum’s standard deviations $\sigma_{J,vr}$ and, if necessary, the stratum’s standard deviations $\sigma_{Z,J,vr}$. We denote as $V$ the overall number of interest variables. It must be noted that these statistics are predictions of the unknown target parameters, or are stratum means and standard deviations of proxy variables known at the population level. It is assumed that these proxies are correlated with the phenomena of interest.

Furthermore, the dataset includes the domain type variables. For example, if it is sought to control the sampling errors at the levels of national, regional and economic activity, the strata can be defined as regions by economic activity and the dataset has three domain-type variables: the national domain (constant) variable; the regional domain variable that identifies the region encompassing the stratum; and the economic activity domain variable that identifies the economic activity defining the stratum.

Finally, the dataset encompasses the variable cost per unit that must be constant in the stratum, and an identifying variable if the stratum is a take-all stratum (its value being equal to 1) or not (its value being equal to 0). The variable’s name in the dataset must be written in capital letters. Table 5.1 shows the structure of the design dataset.
Table 5.1 – Example of dataset with information on the stratified design.

The second dataset includes the precision thresholds in terms of the coefficient of variations for the parameters of estimation at domain level. The dataset records are the types of domain, and the variables are the the domain type identifiers and the coefficient of variation threshold variable for each domain type. Table 5.2 below exemplifies the structure of the second input dataset.

It should be noted that MAUSS-R software uses “.txt” file input datasets.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Domain type identifier</th>
<th>CV% threshold of the first variable</th>
<th>…</th>
<th>CV% threshold of the Vth variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable dataset name</td>
<td>DOM</td>
<td>CV1</td>
<td>…</td>
<td>CVV…</td>
</tr>
<tr>
<td>Values</td>
<td>DOM1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>…</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DOMQ</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2 – Example of dataset with precision thresholds.

To run the software, some steps must be performed: (1) a new project must be started; (2) the folder and the input datasets must be selected; (3) a window to view the parameters for the allocation process must be opened; (4) the parameters may have to be modified; (5) a window to view the precision thresholds must be opened; (6) the precision thresholds may have to be modified; and (7) the procedure must be run.

Figure 5.1 depicts the process of running the procedure. The online help system (shown in the initial window) provides all instructions (in English) necessary for an effective use of the software.
Remark 5.5.1: Installing the Interface in the Windows System (“Installazione.txt” file):

A. Run file setup_MaussR.exe.
B. Set PATH variables.

Example for Windows 7:

Start -> Control panel -> System -> Advance system settings -> Environment Variables …

Select PATH in the System variables section and click the “Edit” button. Add before the path lines of the folder including the java.exe and R.exe files, divided by a semi-colon (";").

Example: PATH=C:\Program Files\Java\jre1.6.0_03\bin;C:\Program Files\R\R-7.1\bin;C:\WINDOWS\system32;C:\WINDOWS;C:\WINDOWS\System32\Wbem.
Figure 5.1 – Example: running the MAUSS-R software for the sample allocation.
5.5.2 Numerical example using the data from Mozambique

The results of this simulation summarize the extensive application proposed by FAO (2014). It is based on an artificial data frame generated by two different sampling frames of Mozambique.

The first population’s sampling frame is the household census database (year 2007), hereinafter denoted as $H$. The database’s records are the individuals (heads of household) involved in agricultural, fishing, or forestry activities. The database’s original dimension is of approximately 54,000 records. It includes several sociodemographic environmental and economic variables. Table 5.3 shows the variables used in the experiments.

The second database gathers the information of the large and medium farms census and of a sample of the survey of small farms (year 2009), hereinafter denoted as $F$. The number of records is of approximately 890 farms, and there are environmental and economic variables (see Table 5.3 below).

The two databases were merged, creating artificial links between individuals and farms. The exercise does not seek to predict the links that actually exist in the two populations, but rather to define a realistic context for the ISF. The merging procedure exploited the following variables: for individuals – the type of job and the district of residence; for farms – the sector, the district and the number of employed persons by type. Before merging, a cleaning step for $H$ was carried out to discard the records that did not feature the job type variable (approximately 9,000 records).

<table>
<thead>
<tr>
<th>Type Variable</th>
<th>Household</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sociodemographic</td>
<td>Gender and age of the head of household, Rural/Urban household</td>
</tr>
<tr>
<td>Environmental</td>
<td>Residence District (7, 8, 9), Enumeration Area</td>
</tr>
<tr>
<td>Economic</td>
<td>Cattle, Pigs and small ruminants, Trees, Fishing (yes/no), Type of Employment (worker in private or public/governmental farm, worker in family farm, farm holder, farm holder without employed persons, etc.)</td>
</tr>
<tr>
<td>Farms</td>
<td></td>
</tr>
<tr>
<td>Environmental</td>
<td>District (7, 8, 9), Enumeration Area</td>
</tr>
<tr>
<td>Economic</td>
<td>Cattle, Pigs and small ruminants, Poultry, Employment by type (number of workers, family workers, etc.) Farm Sector (private or public/governmental)</td>
</tr>
</tbody>
</table>

Table 5.3 – The variables used in the simulation with data from Mozambique
Subsequently, it can be noted that approximately 36,000 records of \( H \) declare being farmholders without any employed persons. For these cases, a one-to-one farm-individual link is defined. The remaining individuals were linked with the 890 farms, according to the following hierarchical rules:

1. when possible, each farm was linked to a number of individuals equal to the number of workers, depending upon employment type;

2. first, individuals and farms in the same district are linked. Several residents of District 9 are linked with farms in the other two districts, because the number of workers of the farms in District 9 is lower than the number of individuals that reside in the same district;

3. Individuals are linked with private/public governmental farms when the type of employment and the farm sector agree.

The individual links are computed (see Box 5.4 in Section 5.3.1).

Two scenarios are compared. The first scenario does not assume an ISF. A sample of farms is directly planned with the MM method and the expected CVs of the indirect sample of individuals are computed. The second scenario assumes a sample allocation that takes into account the direct sample of farms and the indirect sample of individuals.

We consider an SSRS design for the farms. The strata are defined as districts by size class (1, 2, 3-4, 5-9, 10-19, 20-49, 50-99, 100+), thus obtaining 21 strata.

The parameters of interest concern the farms and the households. For the farms, cattle, small ruminants and pigs, and poultry at farm level are considered; for the households, cattle, small ruminants and pigs, trees, and fishing are considered.

In the first scenario (without an ISF), the CV thresholds are fixed only for the farm estimates (Cattle, Pigs and small ruminants, Poultry) at the province and district levels (see Table 5.4, CV1, CV2 and CV3).

<table>
<thead>
<tr>
<th></th>
<th>DOM</th>
<th>CV1</th>
<th>CV2</th>
<th>CV3</th>
<th>CV4</th>
<th>CV5</th>
<th>CV6</th>
<th>CV7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Province</td>
<td>DOM1</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>District</td>
<td>DOM2</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 5.4 – CV thresholds for the sample allocation based on the simulated data from Mozambique
In the second scenario (where an ISF is available), the thresholds for the household estimates (Cattle, Pigs and small ruminants, Trees, Fishing) are fixed at 10% at the province level and 15% at the district level (see Table 5.4, CV4, CV5, CV6 and CV7). The cost per farm is fixed as equal to 1. Graphs 5.1 and 5.2 below show the expected CVs after the sample allocation.

**Graph 5.1** – The expected CV (%) of household province estimates in Scenarios I (without ISF) and II (with ISF).

**Graph 5.2** – Maximum expected CV (%) for household district estimates in Scenarios I (without ISF) and II (with ISF).
The analysis focuses on the performance of the sample allocation with respect to the estimates for households. In the two scenarios seen above, the farm sample size is of approximately 4,000 records. For the first scenario, the expected CV for the farm estimates is naturally lower than the expected CV that may be observed when the allocation is based on Scenario II. This is because in the second allocation, part of the farm sample is required to define the indirect sample of head households (FAO, 2014, studies this inefficiency issue in greater detail). The expected household sample size is of approximately 5,300 records. Graph 5.1 shows that in the two scenarios, both CVs are under the 10% threshold. Nevertheless, in the district domains, the CVs of Scenario I are very high (Graph 5.2), exceeding 30% in two cases. The integrated approach to allocation enables the CVs to be controlled.

5.5.3 Numerical example using the data from Burkina Faso

The second numerical example uses the information from the Burkina Faso database of agricultural households obtained from the General Census of Agriculture (RGA) mentioned in Section 2.2. The data is used to simulate an integrated sample allocation in which households are observed by means of a direct sample and farms are sampled indirectly.

The two populations are not linked, and the auxiliary information on farms can be taken from the estimates of totals or from the known totals of proxy variables correlated by the phenomena of interest. In practice, Burkina Faso’s official statistical body produces farm estimates for the province level.

The RGA database features 1,312,557 agricultural households (records). For each household, auxiliary information is available on the environment, the sociodemographic context, and economic variables relating to agricultural activity. Among these, the number of households living within the site reaches the total of 1,715,749; and the people living within the site are, in total, 11,108,379.

The simulation applies the sample allocation according to Example 5 of Section 5.3 (Box 5.5) above. Therefore, the households are the records of the RGA database and it is assumed, approximately, that $M_h$ represents the people living within the site.

The RGA database includes a farm code variable. In particular, each farm code is joined by one or more records. Then, farm statistics are simulated, summing the values of the Cattle and Sheep & Goats variables for the records that join the same farm identifier. The $Y_{F,v}^A$ at the province level A can thus be found.
### Table 5.5 – Variables used in the simulation with data from Burkina Faso

The numerical example takes into account a sampling design that is similar to the annual agricultural surveys conducted in Burkina Faso on a permanent basis, which use a sample of households selected from the RGA’s sampling frame. Here, a stratified simple random sampling design is considered. In practice, a two-stage sampling plan was adopted. Having defined the household sample sizes, a PPS sample of villages was selected, and 5 households in each village are interviewed.

In the simulation, the same household stratification is performed so that the stratum is the province by Type of agricultural activity for 118 strata; this yields the allocation necessary to control the sampling errors, at both country (Domain type identifier: DOM1) and province levels (Domain type identifier: DOM2) for 45 provinces, as in the current household survey. Different CV thresholds are established for the six variables (Table 5.6 below).

### Table 5.6 – CV thresholds for the sample allocation based on simulated data from Burkina Faso

The first four variables refer to the households (Rain-fed Crops (yes/no), Horticulture (yes/no), Fruit trees (yes/no), Equipped household (yes/no); CV1, CV2, CV3 and CV4), and the fifth and sixth refer to the farms (Cattle, Sheep & Goats, CV5 and CV6). The different CV values reflect the different relative variances between the six variables. The higher the relative variability for a given variable,
the larger will be the sample size to satisfy the threshold; thus, to limit the sample size, it is acceptable to plan higher sampling errors for problematic variables. The cost per household is fixed as being equal to 1.

Finally, the computation of the mean and standard deviation values of the six variables M1 to M6 and S1 to S6 considers the mean and standard deviation observed in the stratum.

Given these inputs, the final allocation produces a sample size of 7,619 agricultural households.

The software provides further results and outputs (in the form of .xls files) that can be viewed in the Population Report and the Sample Allocation Report (respectively, Figures 5.2 and 5.3 below).
References


