INDIRECT SAMPLING, A WAY TO OVERCOME THE WEAKNESS OF THE LISTS IN AGRICULTURAL SURVEY

Piero, Falorsi
ISTAT, Directorate for Methodology and Statistics Process Design
Via Cesare Balbo, 16
Rome, Italy
falorsi@istat.it

Angela, Piersante
FAO, Statistics Division
Via Terme di Caracalla, 1
Rome, Italy
angelapiersante@gmail.com

Dramane, Bako
Global Strategy to improve Rural and Agriculture Statistics, FAO Statistics Division
Via Terme di Caracalla, 1
Rome, Italy
Dramane.Bako@fao.org
DOI: 10.1481/icasVII.2016.f35e

ABSTRACT

The growing demand by policy and decision makers for statistics based on information that is interlinked in economic, social, and environmental aspects, requires a large-scale expansion, in terms of organization and budget, of efforts to implement statistical surveys. In developing countries, agriculture (broadly including fisheries and forestry) is the predominant activity that is interconnected with all these sectors.

Required data to support the development of agricultural projects are usually collected by sector, using different sampling frames and methodologies, without any possibility to measure the cross-sector impact of a given action, consequently affecting the quality of the statistics generated by scattered and manifold data collection methodologies.

Developing a sampling frame for a sector as complex as the rural sector is difficult and expensive for many countries and it is often impossible to establish a frame for certain specific statistical units.
For most countries, the General Census of Agriculture and Population Census, which are usually conducted every ten years, are the only statistical operation that builds lists for agricultural surveys. However, these sources alone cannot provide a sampling frame for all statistical units of interest for data collection in the rural sector, therefore possible lists could be incomplete, unavailable or obsolete.

The *Indirect Sampling* with the application of the *Generalized Weighted Sampling Method (GWSM)* could represent a suitable cost-effective method, capable of offsetting the shortcomings of lists, for rural and agriculture data collection. This method produces estimates for the unknown population object of interest, by calculating the weights of each sampled statistical unit for which there is no list, using the weights of the sampled units of a population for which a sampling frame exists. The essential requirement is the existence of a relationship between the units of the available frame and the units of the target population compared to the phenomenon to be surveyed.

Based on the Indirect Sampling and GWSM, a further extension which observes two populations jointly has been analysed to develop an integrated survey framework that can propose an alternative to the current multipurpose surveys, with the aim of reducing implementation costs as well as producing unbiased estimations and improving the quality of data collection.

Works on the application of the indirect sampling method on agricultural surveys have been developed by the lead of the Research Program of Global Strategy to Improve Agricultural and Rural Statistics.

Keywords: Indirect sampling, Agricultural statistics, Rural statistics, Integrated survey Framework**,** Cost-effective methods, Weight Sharing Method, Link between units of different populations

1. Introduction

Agriculture sector is a vital activity of the developing countries for their economy as well as a mean of subsistence of most rural population. The "growth in the agriculture sector is about two to four times more effective in raising incomes among the poorest compared to other sectors. This is important for 78 per cent of the world's poor who live in rural areas and depend largely on farming to make a living. Agriculture is also crucial to economic growth: it accounts for one-third of gross-domestic product (GDP) and three-quarters of employment in Sub-Saharan Africa."[1]

The statistics demanded by policy and decision makers are based on information that is interlinked in economic, social, and environmental aspects, which require a large-scale expansion of national efforts to implement statistical surveys, in terms of organization and budget. An alternative approach for collecting data, by integrating information from different sources using cost-effective methods, is becoming a crucial requirement for the production of reliable statistics. Especially for developing countries, where the purpose or objectives, for which the information is required, depend on available budgetary resources and time constraints, the combination of various methodologies could be the only possible solution to generate rural and agricultural statistics.

Data collection on agricultural statistics varies by the type of data to be gathered[2] and across countries in terms of local items, periodicity and methods and is carried out by sampling methodologies (agricultural census and surveys) conducted on agricultural holdings. Indeed, other sources, such as population censuses, administrative reports and household sample surveys, though not specifically focused on the entire agricultural sector, may still provide relevant information.
Sample surveys conducted on holdings do not only collect economic data, but they can also be sources of relevant information on social dimension as well as on agricultural practices that affect the environment. 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.

Although the surveys integration can improve data collection by using a single consistent sampling frame to gather data on several domains, field experience has revealed that sometimes the lists of target populations can be outdated or incomplete. The production of the related statistics would require alternative strategies to those based on the classic direct sampling methods. The proposed approach is a method that deals with cases of unknown or rare populations such as the *Indirect Sampling* with the application of the *Generalized Weight Share Method (GWSM)* developed by P. Lavallée.[3]

[1] World Bank. March 2016. http://www.worldbank.org/en/topic/agriculture/overview
[2] Global Strategy, 2015. Integrated Survey framework.
- Current data. These are related 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.
- 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 to be compiled on a frequent basis; compilation every five or ten years is sufficient. These data are usually collected through agricultural censuses.

[3] Lavallée P. (2007), Indirect Sampling, Springer Series in Statistics, ISBN-10:0-387-70778-6

This method sets up a framework that relies mainly on the concept of relation among groups joined together by common characteristics which are identified by clusters and classified by the observational analysis subject to the study, such as economic activities, recipients of services, frequented places, recreational activities during certain periods of time and so forth.
Therefore, a statistical population can be considered as if it comprised sub-sets, which present common features and may knowingly or unknowingly have a cluster structure. Such groups are identified by the phenomenon being studied and are in relation with other statistical populations.

In practice, when a survey is conducted on a target population that is unknown or rare, the relation with the known population is not evident, therefore the path to tackle is to analyse the behavioural observation of statistical units between the target population (which may be rare or hard-to-reach) and another known population where the linkage between the two populations can be identified during the data collection phase. The observation defines the relation that comes out by submitting specific questionnaires. The relation, between the units of the two populations, can be either at individual level or at cluster level and is the essential component of the framework, which determines the number of links needed to calculate the weights of the target population units.

Operationally it means that, if two populations U^A and U^B are related to one another on the specific object of study and only the sampling frame of the population U^A is available, "it is possible to imagine the selection of a sample from U^A and produce an estimate for U^B using the existing links between the two populations. This is what we can refer to as *Indirect Sampling*" (P. Lavallée, 2007).
In due course, the *Generalized Weight Share Method (GWSM)* calculates the weight of each sampled unit of population U^B, using the numbers of its links with the population U^A and the weights of the sampled units of population U^A linked to it.

Furthermore, this method can be extended to develop an integrated survey framework, by observing two populations jointly who are in relation (P. Falorsi, 2014)[4]. The adoption of this framework may be the best solution when dealing with multipurpose surveys such as the following examples:

- sample surveys conducted on households, which do not only collect demographic data, they are also sources of information on the working status and economic well-being. Willing to analyze the contribution of women's work activities to agriculture, it would be interesting to use a households list which includes the entire female population (rural and urban), to extrapolate the second population related to the holdings which could include both female owners and female employers. Otherwise, the latter would have been automatically excluded if the initial list had been that of the holding. An integrated survey should jointly observe and analyze:
  1) the household sector to collect demographic data and,
  2) 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.

- livestock surveys, which provide significant contribution to national income as data is collected to estimate the yields of milk, eggs, meat, feeding as well as the related work-employed, and responding to environment mitigation purposes by gathering information on

[4] Global Strategy, 2014. Technical Report on ISF. Chapter 3. http://gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf

management practices. Unfortunately, the surveys are usually conducted to provide data on food supply (milk, eggs, and meat) separately from specific information gathered on off-season and part-time employment for households and raw material for industries (wool, hides, skins, hair, bristles, etc.), thereby making costly and complex multipurpose surveys. Also in this case, a sample survey framework, that integrates the data collection on holdings and households starting from the holding list, should jointly observe and analyze the two statistical populations, to produce the related estimates.

In conclusion, when the traditional direct sampling strategy is not convenient due to cases of "hard-to-reach" [5] populations or budget constraints, the Indirect Sampling method can then be used because it deals with cases which present the following conditions:

1. No adequate sampling frame as it could be incomplete, obsolete or not available.
2. Possible use of a different frame, due to the relation with the units of target population.
3. A clustering tendency of the target population according to the phenomenon to be analysed.

The advantage of adopting the Indirect Sampling with GWSM is to reduce costs and implementation time, using statistical procedures through the existing and updated frames such as censuses or sample surveys, to estimate unbiased statistics of unknown populations intended as either rare populations or populations with obsolete lists. It is also possible to develop integrated surveys by observing jointly two populations that are in relation, thus improving the data quality based on harmonised statistical units, concept and definitions and classification.

This paper proposes the *Indirect Sampling* with the application of the *GWSM* on some cases in agriculture sector, describing the possible uses on unknown or obsolete lists of the target populations.

2. Indirect Sampling and GWSM applied to Agriculture

Developing countries present a strong preponderant agricultural traditional sector, considered also as subsistence agriculture and attributable to the agricultural households, so that most of the rural areas is employed in this sector. "It is for this reason that in most censuses and surveys, holdings in traditional sectors are identified through household using a list of households" [6], so that a correspondence is often assumed between the Households and traditional Agricultural Holdings [7], thus collecting data on traditional agricultural holdings is equivalent to collecting data on households. "*A farm household can be defined as a household in which any member has both an agricultural main activity and a status of "own account worker*"[8].

Although, the Guidelines for Linking Population and Housing Censuses with the Agricultural Censuses recommends and supports the coordination between the two censuses to be consistent with the reference lists, concepts and definitions, in practice many countries do not have appropriate resources in terms of budget, well-skilled technical personnel, good organizational systems to coordinate the activity of linking the two censuses, as well as keeping updated lists.

[5] Marpsat and Razafindratsima. 2010. Survey methods for hard-to-reach populations: introduction to the special issue. http://mio.sagepub.com/content/5/2/3.1.refshttp://mio.sagepub.com/content/5/2/3.1.refs
[6] FAO.1983. Paper series 35/Prov. Use of household surveys for collection of food and agricultural statistics. Rome
[7] FAO. 2005. WCA 2010, paragraphs 3.27 to 3.35. http://www.fao.org/docrep/009/a0135e/A0135E04.htm - ch3.5
8 FAO. 2012. Guidelines Linking Population and Housing Censuses with Agricultural Censuses http://www.fao.org/docrep/015/i2680e/i2680e00.htm

In addition, the social-economic organization of the traditional agricultural systems, which is based on many relations between households and holdings, makes gathering data of small-scale farmers very difficult at this scattered and manifold level. In fact, the holdings or farm households are clustered to manage activities often carried out separately on crops, livestock and fisheries. The management structure is usually geared towards the three types of subsistence farm organizations: i) one household managing one holding, ii) one household managing more than one holding, iii) more than one household managing one holding[8].
In addition, a further organization is made up of many households, which cooperate for a common purpose in joint activities for many holdings.

In this regard, the indirect sampling, does not represent a cost-effective method only, is also a suitable approach to deal with the complex relations of the management structure based on the correspondence between household and traditional holding, worthwhile to conduct integrated socio-economic surveys by using the most updated list frames, either related to households population or holdings population. Furthermore, it is capable of obtaining most of information from heterogeneous groups of individuals at the same time instead of collecting few data from single individuals spread out in different locations, thus saving time and costs.

Figure 1 below shows the possible links that may exist between the households' population (U^A) and the holdings population (U^B), compared to the above-mentioned agricultural traditional system. In fact, the household members belonging to U^A are the individuals who participate in agricultural activities either directly by themselves or indirectly by their own relatives or workers. The holdings (clusters) belonging to U^B represent the households' economic organization.

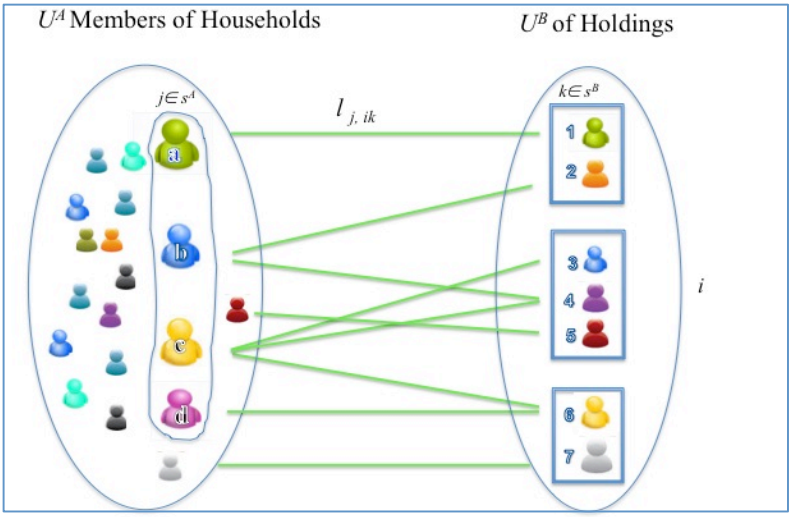

Figure 1: *The relationships between households and holdings*

The observation analysis of the relationships ($l_{j,ik}$) between each member j belonging to the U^A and each individual k of clusters i belonging to U^B, reveals the correspondences between the individuals of the two populations as well as the number of the total links.
This example, also, shows that the links can be one-to-one (case of j = a corresponding to k= 1), one-to-many (case of j = b corresponding to k= 2 and k = 4; or j= c to k= 3, 4 and 6), many-to-one (case of j = c and j=d to k= 6) and many-to-many (case of j = c and j=d to k=3,4,6).

The rectangles represent the holdings (or cluster *i*) of U^B. Each cluster *i* comprises the links and all individuals *k* taking part in agricultural activities, including those not in correspondence with the sampled household members $j \in s^A$ as the cases $k=5$ and $k=7$.

The links are identified between all *j* members of U^A and *k* individuals of each *i* cluster belonging to U^B, with $l_{j,ik} = 1$ if a link exists and with $l_{j,ik} = 0$ in other cases. Each cluster *i* must have at least one link[9] with one sampled individual *j* of U^A that is:

$$L_i^B = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} l_{j,ik} > 0$$

where M_i^B is the size of cluster $i \in U^B$,
M^A is the size of U^A.

Unearthing the links is an exploratory activity on the field to collect reliable information conformed to the survey's objectives, which require the submission of questionnaires as simply and appropriate as possible. A pilot survey and pre-tests could be very useful for analysing possible issues and eventually to work them out.
All individuals within the same cluster $i \in s^B$ (where s^B is the sample of clusters observed in U^B) must be interviewed in order to provide both the measure of the variable of interest y_{ik} and the number of the links $L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$ between *ik* individuals of U^B and *j* household members of U^A.

The final step of the framework building is the weighting procedure of the GWSM to associate the weights to each individual of U^B, by using the weights of the household sampling frame of each members $j \in s^A$.

The sections below are mainly focused on the procedure of computing the weights to produce the estimates of the target population (which may be rare) starting from a known population. It is also shown the process on how to observe two populations jointly.

2.1 Main steps of the method

The following steps describe the process to be carried out, assuming that the list of households' population U^A is available and that the holdings' population U^B is the target population.

1) Selection of the probability sample according to a certain sampling design that could be, for example, a traditional multistage Proportional to Size Stratified design of the household population and calculation of the inclusion probability for each individual to get:
   - Sample s^A that contains m^A individuals selected from U^A of size M^A.
   - Selection probability $\pi_j^A > 0$ of all sampled members $j \in s^A$.

2) Analysis and observation of the existing relationships (links) between populations U^A and U^B, to set up the clusters of the population U^B as follows:
   - Population U^B contains M^B individuals
   - U^B is broken down into N^B clusters, where cluster *i* contains M_i^B individuals.

[9] Lavallée P. (2007), Indirect Sampling, Springer Series in Statistics, ISBN-10:0-387-70778-6

The information on the correspondences and links between the j members of U^A and the k individuals of U^B is gathered by specific questionnaires based on the subject to be surveyed.

For instance, the questionnaires could identify the links by investigating the specific working role of the household members performed in the holdings such as holder, sub-holder, manager, worker, co-worker as well as enhancing further information about other individuals of the holding involved in the agricultural and rural activities. The questions should be submitted by a face-to-face method with a semi-fixed structure to make it possible to pose additional questions (probing) that may not be included in the original format. In this way, getting further information can facilitate the comprehension of the relations, and obtain the interviewer's collaboration, build a certain degree of confidence, verify that the response is appropriate, as well as assist the respondent if the question is difficult to understand. [10]

3) Assignment of a weight to each individual belonging to the holdings population U^B by applying the GWSM. This entails:

- computing the initial weight of each k individual of clusters $i \in s^B$ by the calculation of the weights of sampled j individuals of U^A

$$w'_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} l_{j,ik}$$

where $t_j = 1$ if $j \in s^A$, and 0 otherwise,
$\pi_j^A > 0$ is the inclusion probability of individual $j \in s^A$

- calculating the total number of links for each k individuals of clusters $i \in s^B$

$$L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$$

- computing the final weight w_i of each cluster $i \in U_i^B$

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}^B}$$

where M_i^B is the size of cluster $i \in U^B$

- assigning $w_i = w_{ik}$ for all $k \in U_i^B$:

$$w_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \frac{L_{j,i}}{L_i^B},$$

where $L_{j,i} = \sum_{k=1}^{M_i^B} l_{j,ik}$
$L_i^B = \sum_{k=1}^{M_i^B} L_{ik}^B$ that is the total number of the links of cluster i.

[10] Global Strategy. 2015. "Guidelines of Integrated Surveys Framework" Chapter 3 part I, page 58.

4) Estimation of the total of the variable $\hat{Y}_B$ related to the holding population is calculated with the Horvitz-Thompson estimator using the weights computed by GWSM as follows:

$$\hat{Y}_B = \sum_{i=1}^{n^B} \sum_{k=1}^{M_i^B} w_{ik} y_{ik}$$

where n^B is the number of surveyed clusters and w_{ik} is the weight assigned to j individual belonging to i^{th} cluster.

The links may be defined at unit level or at cluster level. This is illustrated in figure 2 below where, in the case of unit level (the left hand side of the figure), the index j refers to a single individual. In this situation, the total number of links of cluster F1 is equal to 4. The case of links defined at cluster level is described on the right hand side of the figure where the index j refers to a household. The total number of links of cluster F1 is equal to 3.

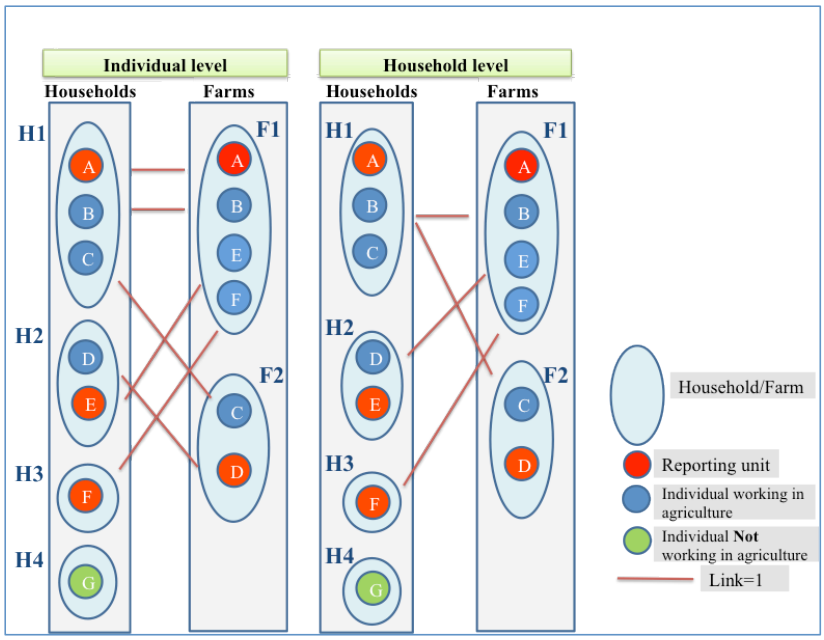

Figure 2: *Example of the links between individuals or between clusters*

2.2 Application to two populations jointly observed

There are cases where it is necessary to undertake surveys that require crossover analyses such as, for example, a study concerning the contribution of women's employment to agriculture in developing countries. In this case, the access to resources control must be gone through, 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.

The joint observation of households and farmers 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.

The scheme consists of a two-step procedure and depends on the available frame, such as:

- if the households list is available
  a. the households are observed in the first step and based on direct sampling,
  b. the farms are observed in the second step and based on indirect sampling.

- if the farms list is available
  a. the farms are observed, in the first step,
  b. the households are observed, in the second step.

The following example shows the case of having the availability of the households' list by selecting the sample with standard stratified multistage sampling design.

The sampling units are stratified according to geographical criteria, where the Enumeration Areas (EAs) of the Census are the Primary Sampling Units (PSU) and all households, belonging to sampled EAs, are observed as Secondary Sampling Units (SSU) with the respective inclusion probability.
In this regard, suppose that the households H1 and H3 were selected from two different EAs and hypothesize that the related inclusion probabilities were respectively equal to $\pi_{\mathrm{H1}} = 0.001$ and $\pi_{\mathrm{H3}} 0.02$ (it would mean that H1 has been selected from a set of 1000 units and H3 has been selected from a set of 50 units).[11]

At a later stage, in which the households are observed in the first step (the left hand side of figure 2), the links are defined with the support of two questionnaires to get the following result:

Step 1 The enumerator starts interviewing the first Household, the head of H1 to collect data on gender and gather information on the farm's links. It is clear that the Household is linked to Farms F1 (with two links) and F2 (with one link).

The second interview is submitted to the head of Household H3 to collect the data on gender and information on the farm's links with the same method adopted for Household H1.

Step 2 The enumerator interviews the agricultural holder of Farm F1 and collects the data on the parcel/plot tenure. Then, he gathers the information on the farm's total links which in this case are equal to 4, where

- the individuals A, B and F represent the links with the members of sampled households H1 and H3,
- the individual E which is linked to a member of H2, therefore it cannot be considered a link because H2 is not sampled.

[11] Global Strategy. 2015. ISF- Chapter 2.

The GWSM weight of farm F1 may then be computed as reported on figure 3:

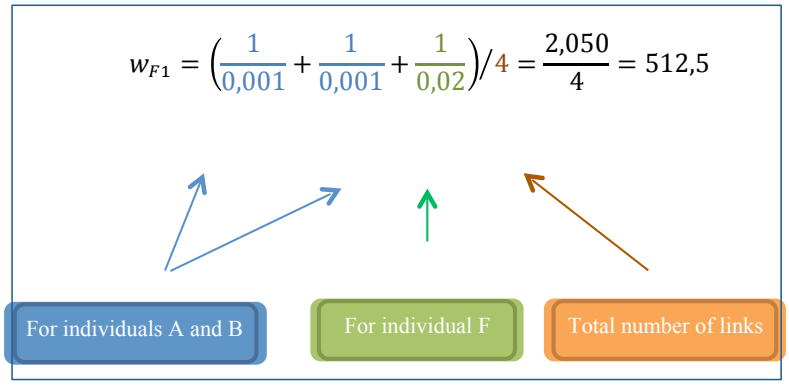

Figure 3: *Calculation of the final weight*

In the due course, the second interview is submitted to the agricultural holder of Farm F2 to collect data on the parcel/plot tenure and information on the farm's links, following the same procedure adopted for Farm F1.

Estimation

Let Y^A 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, whose list is available; and let Y^B be the total of interest (e.g. the total land area in the country) referring to the population of the farms, observed indirectly.

Let s^A be the sample of observed households and s^B the sample of observed farms.

The sample estimate $\hat{Y}^A$ of the total Y^A is obtained by:

$$\hat{Y}^A = \sum_{j=1}^{s^A} y_j \, w_j$$

where y_j is the value of the variable of interest of the individual j (e.g. $y_j = 1$ if the women j within the household performs agricultural activities) and w_j is the sample design weight.

Likewise, the sample estimate $\hat{Y}^B$ of the total Y^B is simply obtained by assigning to each farm i observed in s^B a weight w_i obtained with the weight share method (see section 2.1):

$$\hat{Y}^B = \sum_{i=1}^{s^B} y_i \, w_i$$

where y_i is the value of the variable of interest (e.g. the total parcel/plot land tenure in the farm) of the farm i and w_i is the weight calculated by GWSM.

3. Case studies

The cases described below have been taken from the recent research program of the Global Strategy to Improve Agriculture and Rural Statistics (www.gsars.org.) which deals with methodologies to collect data in order to make them comparable cross over the countries, respecting the international quality standards as well as to facilitate the development of cost effective methods.

The reported examples refer to some studies to apply the indirect sampling with GWSM taken from the "Guidelines of Integrated Survey Framework" and the "Guidelines on the Enumeration of Nomadic and Semi-Nomadic Livestock" to simulate the practical use on how to build a logical framework, based on articulated questionnaire[12], as well as the running of weights calculation. They can represent some opportunities to develop integrated surveys and improve initiatives on the field to adapt this method as needed.

The following sections describe three proposals of this approach for the use in agriculture sector in:

- estimating statistics on holdings starting from a household frame,
- updating of the units weights in case of change in the statistical units,
- producing estimates for the Nomadic and Semi-Nomadic Livestock.

Detailed information on these subjects is available on the web site of the Global Strategy.

3.1 Estimating statistics on holdings starting from a household frame

An interesting application is the case carried out by Burkina Faso to conduct a national study to collect information on agriculture sites where several farmers work individually or in groups for the production of rice, corn, vegetables etc. Specifically, this study should produce estimates on the area harvested, crops production, number of farmers and farmers' incomes working in these sites. In particular, irrigated crop production is practiced mainly on a number of sites, developed by the State's funds as well as by Non Governmental Organizations (NGOs) and other private projects.

The economic-social structure is organized to provide each farm's individual with the sites and the parcels, however the limited number of developed sites does not allow several producers to obtain a parcel for irrigated production on a developed site. Thus, for many households, their links to irrigated sites are limited to the work of some of their members as permanent or temporary employees on these sites and members of the same household can work on different farms. At the same time, 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. Therefore, for the objective of this survey it would be necessary to include all these employees in order to cover a larger number of sites.

Moreover, Burkina Faso conducts permanent annual agricultural surveys based on the "Recensement General de l'Agriculture" (RGA) sampling frame and as the purpose of the matter, here described, refers to the farm household, intended as farm household of traditional agricultural systems, the known population is represented by the available household sample frame selected from the RGA frame and used in other surveys. This sample appears to suit to our study also for the following two reasons:

[12] Global Strategy. 2015. ISF- Chapter 3. Modules and operational rules for observing farms starting from households http://gsars.org/wp-content/uploads/2015/05/ISF-Guidelines_12_05_2015-WEB.pdf

- Such a sample will enable the subsequent availability of more information to analyze 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.

Hence, this case is featured by the initial requirements of the indirect sampling method, within the following conditions:

- The farm workers at work sites represent the unknown population.
- The farm household members frame is available to create the links framework.
- Each work site is a cluster of the unknown population.

The following steps describe the application of the above-mentioned approach.

1) Selection of the farm household members sample by a certain design sampling.
2) Analysis and observation of the relationships between the farm household members and farm workers at work sites, defining *a household linked to an irrigated production site if at least one of its members works on this site as a farmer or employee.*

   The links between the household members and the farm workers are built with the two steps procedure using respectively two questionnaires and interviewing:

   Step 1 Each household member, for collecting a great deal of possible reliable information to identify the location of the sites, its total number as well as the working status of the household member as farm owner or employee.

Household Code:	Economically Active Member code:……
Are you a farmer on an irrigated site?	
If yes, list the sites concerned:	
• Name of Site 1.....	
• Name of Site 2.....	
Did you work as an employee on an irrigated site last year?	
If yes, list the sites concerned:	
• Name of Site 1.....	
• Name of Site 2.....	

Table 1: *Questionnaire to observe households and farm sites links*

Once the questionnaires are filled in, the enumerator should complete the list of sites linked to the household, and summarize which sites are linked to the household member. A unique code (such as geo-referential coordinates) is also assigned to each site that is linked to the respondent household member.

	Site Name	Member Respondent Code	Site Code
1.	.....	.....	.....
2.	.....	.....	.....

Table 2: *Summary table on links information from households*

Step 2 Each worker at work sites, for gathering information on farmers and employees, which should incorporate the same site code of the households' questionnaires submitted for the preliminary identification of the links.

Site Name:	
Site Code:	
Name of the Respondent on the Site:	
Questions to the Respondent	• In total, how many farmers are working on this site? .....
	• Can you provide an estimate of the total number of employees working on the site? .....

Table 3: *Questionnaire carried out on the sites*

The result of the above mentioned questionnaires is shown in the framework on figure 4, where the links $l_{j,ik}$ are drawn up between all farm household members belonging to M^A and those individuals k working in each i site belonging to M_i^B (size of each cluster). The inclusion probabilities π_j^A, of each household member $j \in s^A$, compute the initial weights of each individual $ik \in M_i^B$.

Moreover, throughout the survey, data collected in each site *i* have also furthered the list of the individuals to be surveyed, increasing significantly the initial information obtained from the farm household members.

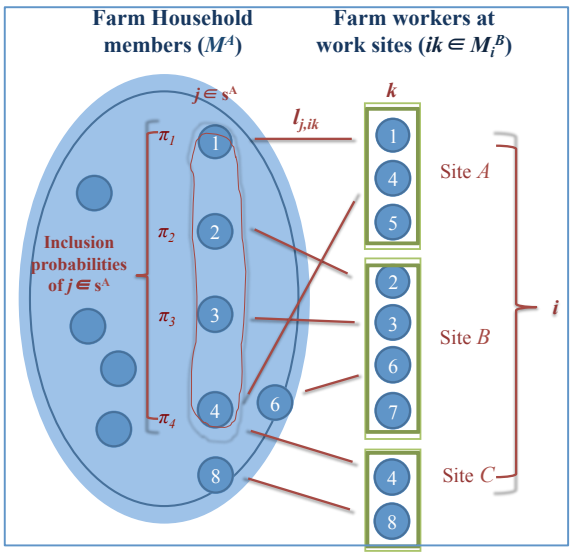

Figure 4: *Correspondences between Farm Household members and Farm workers*

3) Assignment of a weight to each farmer site by applying the GWSM as follows:

- Computing the initial weight w'_{ik} of k individuals in each i site. It is calculated[13] by summing the weights $\frac{1}{\pi_j^A}$ of farm household members $j \in M^A$ linked to each ik, with $t_j = 1$ if $j \in s^A$ (cases 1, 2, 3 and 4) and 0 otherwise (cases 6 and 8). An initial weight zero is assigned to all individuals not having a link (cases 5 and 7).

i	k	$w'_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} l_{j,ik}$
A	1	46
	4	58
	5	0
B	2	65
	3	28
	6	0
	7	0
C	4	58
	8	0

Table 4: *Initial weight estimation of farm workers*

- Calculation of the total number of links in each work site cluster i (quantity L_i^B), by summing for each work site i all links between the farm household members of M^A, both sampled and not sampled ones (quantity L_{ik}^B), and the individuals k of work sites i belonging to M_i^B.

i	k	$w'_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} l_{j,ik}$	$L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$	$L_i^B = \sum_{k=1}^{M_i^B} L_{ik}^B$
A	1	46	1	2
	4	58	1	
	5	0	0	
B	2	65	1	3
	3	28	1	
	6	0	1	
	7	0	0	
C	4	58	1	2
	8	0	1	

Table 5: *Calculation of total number of the links within the work sites*

- Calculation of the final weight w_i for each work site cluster i belonging to M_i^B, by dividing the initial weight with the total links number of each work site.

i	k	$w'_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} l_{j,ik}$	$L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$	$L_i^B = \sum_{k=1}^{M_i^B} L_{ik}^B$	$\sum_{k=1}^{M_i^B} w'_{ik}$	$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}^B}$
A	1	46	1	2	104	104/2 = 52
	4	58	1			
	5	0	0			
B	2	65	1	3	93	93/3 = 31
	3	28	1			

[13] These figures have been made up to run the simulation.

	6	0	0			
	7	0	0			
C	4	58	1	2	58	58/2 = 29
	8	0				

Table 6: *Final weight of each work site*

- Assignment of the final weight w_{ik} to each farm worker *ik* within the work site cluster *i*. These final weights are used to estimate the variable to be measured $\hat{Y}^B$, calculated by GWSM.

i	k	$w_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \frac{L_{j,i}}{L_i^B}$
A	1	52
	4	52
	5	52
B	2	31
	3	31
	6	31
	7	31
C	4	29
	8	29

Table 7: *Final weight estimation of each farm worker at work sites*

3.2 Updating of the units weights in case of change in the statistical units

This case, extracted from a more extensive discussion in the "Guidelines of Integrated Survey Framework"[14] (ISF), deals with the changes of the statistical units of surveys based on ~~a~~ panel agricultural holdings, mainly caused by disappearance, fusion, division and population change. This phenomenon may undergo significant modifications from one year to another, due to endogenous or exogenous events. Some solutions are presented in the ISF Guidelines[15], and among these the indirect sampling with the GWSM has been proposed as an alternative solution to handle the lists updating. The practice here reported refers to the *Changes in households* with the 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.

Although the current solutions are usually sufficient to solve problems related to changes in the composition of households, for cross-sectional estimates in a given year, the system may be interested in the activities, pursued by individual household members, which are relevant to agricultural statistics. In this case, each member becomes a statistical unit, and the weight of a member corresponds to his household weight; indeed, usually all members of a sampled

[14] Global Strategy. 2015. Guidelines of Integrated Survey Framework. http://gsars.org/wp-content/uploads/2015/05/ISF-Guidelines_12_05_2015-WEB.pdf

[15] Alternative methods of calculating the weights are available in the literature. 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).

household are 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 shall be taken into account. 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 the conventional approach because, for example, several members of different households may combine to form a new household, which includes members of other non-sampled households. The weights of all the members of the new household become difficult to estimate.

Also in case of change in the statistical units, the Indirect Sampling with GWSM concerns 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 enable estimation of the weights of the individuals who were within the original sample.

In fact, the current new structure presents the following conditions:

i) the unknown weights of the household members of Wave 2,
ii) the availability of Wave 1 list to be linked to Wave 2,
iii) the members of Wave 2 are clustered by the new households structure.

As reported in the methodological observation section, the links between the two Waves are built with the support of questionnaires to collect information, by interviewing:

Step 1 Households of Wave 1, to make sure that all cluster Wave 1 Households are linked to the new structure of Wave 2 Households.

Household Code:		
Full name of head of household:		
How many members left your household between last year and this year?.....		
Among these members, how many have joined other households?		
Can you name these members and households they joined?		
Member Code	The new Household Code	Name of the head of the new household
.....	.....	.....
.....	.....	.....

Table 8: *Questionnaire to observe the links of Wave 1 Households*

Step 2 Households of Wave 2, to identify the total number of links with households Wave 1, that can be performed at Individual or Cluster level. In case of Wave 2 Household A (figure 4), the total number of the links is 1 because only individual A1 belongs to Wave 1, while individual E is a new entry.
An example of the questions to be posed at individual or household (cluster) level approach can be adopted as Table 9 shows below.

Type of links	Questions
Household level	1. Are there members of your household that belonged to other households last year? (Yes/No) .....

	2. In how many households did these members live? .....
Individual level	1. How many members of your household belonged to other households last year? .....

Table 9: *Questionnaire to identify the links of the two Household Waves*

The observation reveled some changes of the population panel structure between Year 1 and Year 2, Wave 1 turned into Wave 2, as figure 5 shows below:

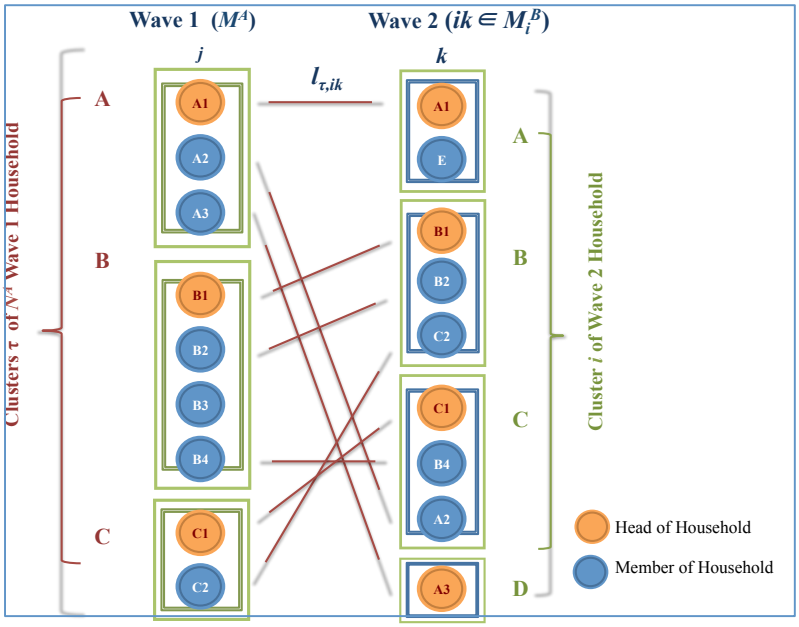

Figure 5: *Framework of changes in the composition of households*

At a later stage, the weighting procedure entails the following steps:

- Estimation of the weights $\frac{1}{\pi_\tau^A}$, for each Wave 1 of N^A cluster τ belonging to M^A, is the inverse of the inclusion probabilities π_τ^A, as Table 10 shows.

Wave 1		
Household (τ *clusters* N^A)	Household Members (j)	$\frac{1}{\pi_\tau^A}$
A	A1	57
A	A2	
A	A3	
B	B1	104
B	B2	
B	B3	
B	B4	
C	C1	63
C	C2	

Table 10: *Structure and weights of Wave 1*

- Calculation of the final weight w_{ik} of each member belonging to Wave 2. Table 11 shows the new structure of the three households A, B, C including the additional D, the initial weights w'_{ik} of each Wave 1 member as well as the total links L_i^B between the two Waves.

Wave 2					
Households (i *clusters* M^B)	Household Members (ik)	$w'_{ik} = \sum_{j=1}^{M^A} \frac{t_\tau}{\pi_\tau^A} l_{\tau,ik}$	$L_i^B = \sum_{\tau=1}^{N^A} L_{\tau,i}$	$\sum_{k=1}^{M_i^B} w'_{ik}$	$w_{ik} = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}^B}$
A	A1	57	1	57	57/1=57
A	E	0		57	57/1=57
B	B1	104	3	104+104+63	271/3= 90,33
B	B2	104	3	104+104+63	271/3= 90,33
B	C2	63	3	104+104+63	271/3= 90,33
B	B3	-	-	-	-
C	C1	63	3	63+104+57	224/3=74,66
C	B4	104	3	63+104+57	224/3=74,66
C	A2	57	3	63+104+57	224/3=74,66
D	A3	57	1	57	57/1=57

Table 11: *GWSM final weights of Wave 2 Households*

3.3 Application of Indirect sampling and GWSM to produce estimates for the Nomadic and Semi-Nomadic Livestock

The nomadic and transhumant livestock[16] enumeration is an important component of the estimate of the total national number of breeding animals, especially for those countries affected by this phenomenon. However, the difficulty and even sometimes the impossibility to collect data through the agricultural census, because of the herders' seasonal or random movements, intra country and across the countries, make the implementation of this activity considerably burdensome both on operational and budgetary aspects.

Among several approaches to data collection, those that are related to the sample surveys on ground, usually collect the information on herds by a questionnaire. It is submitted to the herders when they bring the herds to the Enumeration Points (EPs) or drinking points during the survey year (Lakes, Rivers, Ponds, Wells, Boreholes, etc.)[17]. Although, the EPs are essential sources to collect data, most of them are difficult to access so that the available lists are not often

[16] Global Strategy. 2016. Guidelines on the Enumeration of Nomadic and Semi-Nomadic Livestock, definitions of:
a. Transhumant livestock/pastoralists as not permanently settled; movements characterised by regular, cyclical, short distance movements; livelihoods depend largely on livestock
b. Nomadic livestock/pastoralists as not permanently settled; movements characterised by irregular, erratic, long distance movements; livelihoods depend almost entirely on livestock.

[17] Global Strategy. 2016. Guidelines on the Enumeration of Nomadic and Semi-Nomadic (Transhumant) Livestock. "The identification and definition of enumeration points must be made in a participatory way with all stakeholders: government, local authorities, herders' organizations, and civil society. It is also important to note that the transhumant livestock and nomadic livestock do not always have the same enumeration points. For example in Niger, the enumeration points identified were stock routes for nomadic livestock and water points for transhumant livestock".

completely exhaustive to cover the enumeration list for the entire geographical zone, thus creating many issues in obtaining the needed information.

Also this case could be dealt with the "hard-to-reach" population methods by applying the Indirect Sampling and GWSM to carry out a sample survey on ground. The purpose is to estimate the total number of herds concerning nomadic and transhumant livestock at sub-national level (province), by increasing the initial list of EPs to develop a wider exploratory inquiry on the field. This case is featured by the following conditions:

- The nomadic and transhumant herds are the unknown or rare population, which is in relation with a known population of EPs, where animals gather and are given water, so that it is possible to create a link between the two populations.
- The initial sampling frame of a known population could be made available by selecting a two-stage sampling design at sub national level:
  - the Primary Sampling Units (PSU) as the Enumeration Areas (EAs) of the census (however, it is also possible to select counties, villages or other relatively small geographical areas). The PSUs can be selected either with probabilities proportional to their sizes (which could be the number of EPs in each EA) or equal probabilities.
  - the Secondary Sampling Units (SSU), that correspond to the EPs, are selected for each PSU with equal probabilities in each stratum.
- The herds (the unknown population) can be considered as clusters which group the EPs selected by herder, who gets used to taking the animals to drink.

The observational approach is described by analysing the links between the initial sampled EPs and the real EPs frequented by the herds through the information received by the herders, defining "*a herd as linked to an enumeration point if it frequents that enumeration point at least one time during a period of time (generally a year or several months in a year)*". [18]

Table 12 below proposes a kind of the questionnaire that could be submitted to three hypothetical herders who are responsible respectively for the herds A, B and C. The interview should take place at one of the sampled EPs and should also be widened as much as possible on eventual insights to better gather all needed information that helps in drawing up the framework of the links.

Questions [19]	**Answers**
Q1: How many EPs will you likely frequent with your herd this year?	*A1.* Three EPs by Herd A. *A1.* Four EPs by Herd B. *A1.* One EP by Herd C.
Q2: What are they? (Checking if the declared EP are present in the sample or are additional)	*A2.*Three sampled EPs (1, 2 and 3) are frequented by Herd A. *A2.*Two sampled EPs (1 and 2) are frequented and two not sampled (6 and 7) by Herd B. *A2.* One sampled EP (2) is frequented by Herd C.
Q3: Do you happen to meet other herds in the EPs that you frequent?	*A3.*No
Q4: If yes, how often? And which ones	n.a.

[18] Global Strategy. 2016. Guidelines on the Enumeration of Nomadic and Semi-Nomadic (Transhumant) Livestock.
[19] In case the enumeration points are stratified, these questions must be asked for each stratum.

Table 12: *Example of a questionnaire to build the links*

Figure 6 shows the framework, resulting from the above-mentioned questionnaire submitted to the three herders in the three different EPs, with the correspondences between the EPs frequented by Herd A, B and C and the sampled EPs during the period of the survey. The inclusion probabilities π_j^A of the sampled EPs (1, 2 and 3) belonging to the population M^A are calculated by the sampling design and are used to estimate the weights $\frac{1}{\pi_j^A}$ of each EPs frequented by the Herds.

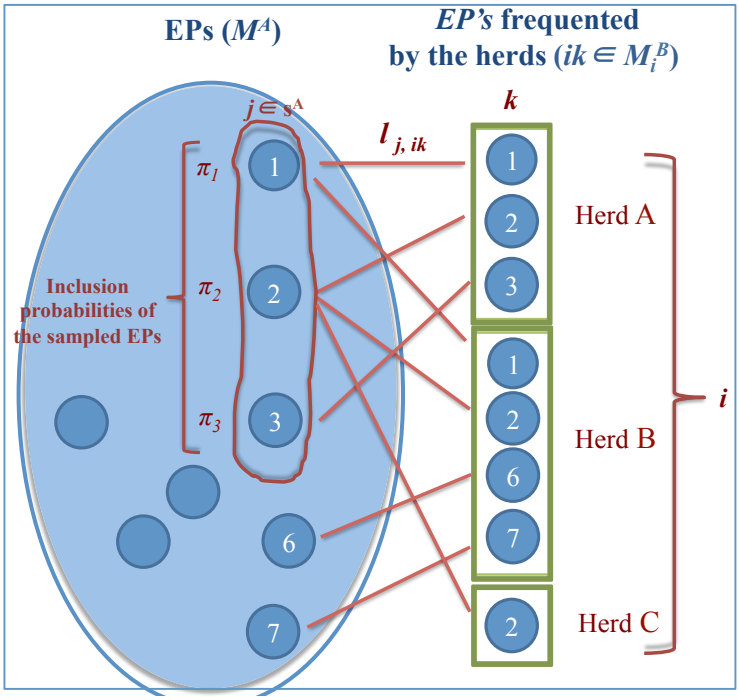

Figure 6: *Correspondences of the EPs and those frequented by Herds*

The following steps describe the application of the weighting procedure with GWSM:

- Computing the initial[20] weight w'_{ik} of the EPs clustered by each Herd. It is the sum of the weights $\frac{1}{\pi_j^A}$ of all the EPs of M^A linked to those frequented by ik Herds, with $t_j = 1$ if $j \in s^A$, and 0 otherwise.

i	k	$w'_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} l_{j,ik}$
A	1	45
	2	28
	3	55
B	1	45
	2	28
	6	0
	7	0
C	2	28

[20] These figures have been made up to run the simulation.

Table 13: *Initial weight estimation of EPs frequented by Herds*

- Calculation of the total number of links for each Herd cluster *i* (quantity L_i^B), by summing for each Herd *i* all links between all EPs (M^A), both sampled and not sampled ones (quantity L_{ik}^B), and EPs clustered by the *ik* Herds belonging to M_i^B (size of the each cluster *i*).

i	k	$w'_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} l_{j,ik}$	$L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$	$L_i^B = \sum_{k=1}^{M_i^B} L_{ik}^B$
A	1	45	1	3
	2	28	1	
	3	55	1	
B	1	45	1	4
	2	28	1	
	6	0	1	
	7	0	1	
C	2	28	1	1

Table 14: *Calculation of the links*

- Estimation of the final weight w_i of each Herd cluster *i***,** that is calculated by dividing the initial weights by total number of the links in each Herd cluster.

i	k	$w'_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} l_{j,ik}$	$L_i^B = \sum_{k=1}^{M_i^B} L_{ik}^B$	$\sum_{k=1}^{M_i^B} w'_{ik}$	$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}^B}$
A	1	45	3	128	128/3 = 42,7
	2	28			
	3	55			
B	1	45	4	73	73/4 = 18,2
	2	28			
	6	0			
	7	0			
C	2	28	1	28	28/1 = 28

Table 15: *Final weight calculation of the clustered EPs frequented by Herds*

- Assignment of the final weight w_{ik} to each *ik* EP for the estimation of the variable subject to the study calculated by GWSM with the estimator $\hat{Y}^B$ above mentioned.

i	k	$w'_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} l_{j,ik}$	$w_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \frac{L_{j,i}}{L_i^B}$
A	1	45	42,7
	2	28	42,7
	3	55	42,7
B	1	45	18,2
	2	28	18,2
	6	0	18,2
	7	0	18,2
C	3	28	28,0

Table 16: *Assignment of the final weight to each EP frequented by Herds*

4. Conclusions

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 exhaustive surveys that can provide reliable estimates at national level.

Furthermore the rural sector is an especially vast and complex sector, where it is not possible to collect all data on the sector during a general agricultural census. For this reason, usually only information on their links with farms is collected. However, data for each of these structures are not collected, although this information is relevant to national statistical program and policy makers.

This is also true for some information that is indirectly related to agriculture, such as its links with the sectors of education, health and infrastructure.

Indirect sampling and the application of GWSM could help 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.

These methods are very useful tools to enable the collection of information using alternative and cost-effective methods that facilitate the data gathering in case of difficulties to get information about rare and unknown populations using statistical procedures.

This paper has presented some examples applied in the agricultural sector that should be deepened and tested on the field at local level in order to meet national needs and eventually develop experiences to strengthen the statistical capacity building of the countries.

REFERENCES

FAO (2010), *Global Strategy to Improve Agricultural and Rural Statistics*, Report Number 56719-GLB, http://www.fao.org/docrep/015/am082e/am082e00.pdf. Accessed on August 12th, 2016.

FAO (2012), *Guidelines for Linking Population and Housing Censuses with Agricultural Censuses with selected country practices*, Special Issue of the FAO Statistical Development Series, ISSN 1014-3378, ISBN 978-92-5-107192-2, http://www.fao.org/docrep/015/i2680e/i2680e.pdf. Accessed on August 12th, 2016.

FAO, Global Strategy (2015). *Guidelines of Integrated Survey Framework.* http://gsars.org/wp-content/uploads/2015/05/ISF-Guidelines_12_05_2015-WEB.pdf. Accessed on August 12th, 2016.

FAO (2005), WCA 2010, http://www.fao.org/docrep/009/a0135e/A0135E04.htm - ch3.5. Accessed on August 12th, 2016.

FAO, Global Strategy (2014). *Technical report on the Integrated Survey* Framework, GO-02-2014, http://gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf. Accessed on August 12th, 2016.

FAO, Global Strategy (2016), *Guidelines on the Enumeration of Nomadic and Semi-Nomadic Livestock. http://gsars.org/wp-content/uploads/2016/08/Guidelines-for-the-Enumeration-of-Nomadic-Livestock-05.pdf.* Accessed on August 12th, 2016.

FAO (1983), Paper series 35/Prov. *Use of household surveys for collection of food and agricultural statistics.* Rome.

Lavallée P. (2007), *Indirect Sampling*, Springer Series in Statistics, Ottawa Canada. ISBN-10:0-387-70778-6.

Marpsat and Razafindratsima (2010), *Survey methods for hard-to-reach populations: introduction to the special issue.* *http://mio.sagepub.com/content/5/2/3.1.refshttp://mio.sagepub.com/content/5/2/3.1.refs.* Accessed on August 12th, 2016.

M. aia (2009), School of Economics & Management, Catholic University of Portugal. *Indirect Sampling in Context of Multiple Frames* *https://www.amstat.org/sections/srms/proceedings/y2009/Files/303803.pdf.* Accessed on August 12th, 2016.