Sampling design for agricultural surveys

Introduction

FAO is recognized by the UN Statistics Division as the organization responsible for statistical standards for a number of sectors including agriculture, forestry and fisheries. The organization has been supporting countries since 1950 to implement their national agricultural census through the World Programme for the Census of Agriculture (WCA) promoting the use of standard international concepts, definitions and methodology.

The census of agriculture is generally aimed to be carried out every ten years while more frequent and quality agricultural data are required for effective planning, financing, and implementation of agricultural development policies, especially in developing countries. Agricultural surveys are naturally a cost effective solution for more detailed and updated agricultural statistics in countries to fulfil both national and international data needs. FAO implemented methodology works on survey design to support technically countries in the implementation of agricultural surveys.

In most cases, agricultural surveys cover both crop and livestock productions but in some contexts specific livestock surveys are required especially for non-sedentary livestock. In addition, for cost effectiveness and data integration in countries, FAO is recommending integrated agricultural surveys covering also fisheries and aquaculture and producing reliable statistics on both households and agricultural holdings. Accordingly, this section will cover broadly sampling designs for (i) agricultural surveys, (ii) livestock (sedentary and mobile) survey and (iii) integrated agricultural and household surveys.

1 Sampling guidelines for agricultural surveys

1.1 Populations

The target population of agricultural surveys consists of agricultural holdings defined by the WCA (FAO, 2015a) as follows: “economic units 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”.

The WCA distinguishes two types of agricultural holdings: (i) holdings in the household sector (operated by households) and (ii) holdings in the non-household sector (operated by other structures like corporations and government institutions). The term “agricultural households” is often used in the household sector to design households that operate agricultural holdings for their own account (either for sale or for own use).

1.2 Sampling frames

For agricultural surveys, FAO recommends the use of master sampling frames for cost-effectiveness, consistency and integration of agricultural statistics in countries. A master sampling frame is a frame that enables selection of different samples (including from different sampling designs) for specific purposes: agricultural surveys, household surveys, and farm management surveys. Such frame enables samples to be

drawn for several different surveys or different rounds of the same survey, which makes it possible to avoid building an ad hoc frame for each survey (FAO, 2015b). Broadly FAO recommends two types of master sampling frames for integrated agricultural surveys (FAO, 2017):

1) ***A multiple frame consisting in two list frames: lists of agricultural holdings (i) in the household sector and (ii) in the non-household sector.***

These two lists could be easily established from an agricultural census.

For the household sector, a cost-effective approach is to link the population and agricultural censuses as suggested by the WCA 2020 (FAO, 2015a). The basic process could be the following:

i. During the preparation of the population census, the Enumeration Areas' geographical limits should be digitally mapped;
ii. An agricultural module to be collected during the population census should be developed, taking into account basic information to stratify the holdings;
iii. After the population census, a complete list of agricultural holdings may be established, together with a complementary list of holdings of the non-household sector, as an MSF for AGRIS. However, in this regard, it is important to note that the processing of the population census data processing may require much time before the data can be used for sampling purposes. The use of CAPI software for data collection will reduce the data processing time.

Alternatively, database from a previous population census could be used if it contains enough information to identify individuals practicing agricultural activities for their own account.

For holdings in the non-household sector, a starting point is using business registers of farms including the national business register and informal business registers of farmers' organisations and making efforts to handle the probable large overlap between them. In addition, all other relevant registers should be considered including the list of government institutions (agricultural research centres, schools, hospitals, prisons etc.) and non-government organisations operating farms. Local knowledge and information from extension agents and local authorities generally also help in that process.

2) ***A multiple frame consisting in an area frame and two list frames (landless holdings raising livestock and large commercial agricultural holdings).***

This is recommended for countries using an area frame for agricultural surveys. Considering that an area frame does not cover landless holdings that raise livestock, a complementary listing of these holdings is recommended. In addition, if large commercial agricultural holdings happen to be sampled from an area frame, they may behave like outliers. A second list of large commercial agricultural holdings is therefore recommended (FAO, 2017).

1.3 Stratification

A primary stratification recommended consists in dividing the main domain's territory into three strata: urban areas, peri-urban areas and rural areas. The limits between these strata must appear on the cartographic material. If available, additional strata based on Agro-Ecological Zones may be relevant, especially to integrate soil fertility and environmental issues. Secondary strata could be based on the essential characteristics of the holdings that may be found in the list frame. Examples of essential characteristics include crop intensity or presence of livestock on the holding. If specific categories of holdings are of particular interest to the country, the use of specific holding-based strata may be considered for later analyses. Administrative data may be helpful for stratification purposes.

In the framework of a two-stage sampling design, the stratification of primary sampling units (PSUs) is highly recommended, because in this type of design, a great part of the overall variability derives from the PSUs' inter-variance. The stratification recommended above should be performed in the sampling frame of the PSUs.

1.4 Sampling designs

The two subpopulations of agricultural holdings (in household and non-household sectors) have generally different characteristics that usually require different sampling designs. The general recommendations of sampling designs are presented in the table below.

Type of frame	**Sub-frame**	**Sampling units**	**Sampling Design**	**Sampling scheme**
Multiple frame of list frames	Holdings of the household sector	AH of the household sector	Stratified two-stage	1st stage: PPS of PSUs (EAs). 2nd stage: SRSWOR of agricultural households
	Holdings of the non-household sector	AH of the non-household sector	Stratified one-stage	SRSWOR of agricultural holdings
Multiple frame of area frame and list frame	Area frame	Segments or points	Stratified two-stage	1st stage: PPS of PSUs (segments or grids/clusters of points) 2nd stage: SRSWOR of points
	Lists	- Landless AH raising livestock - Large commercial holdings	Stratified one-stage	SRSWOR of agricultural holdings

1.5 Longitudinal design

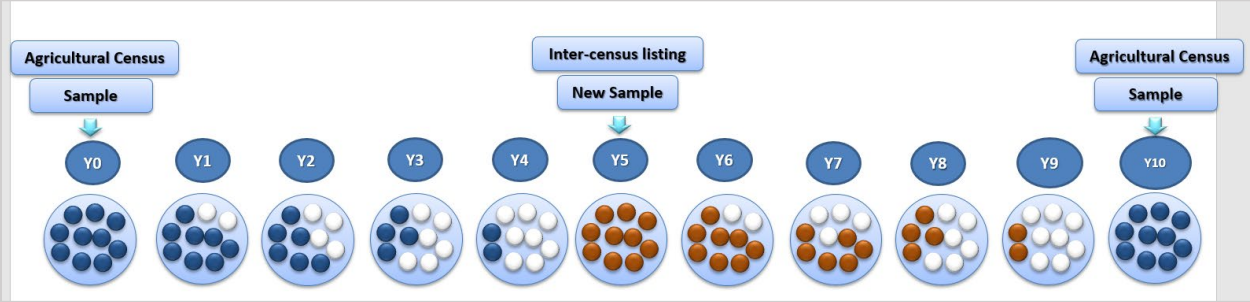
The World Programme for the Census of Agriculture recommends to implement at least one Census of Agriculture every ten years. The census data is then used to build a sampling frame and design a system of agricultural surveys to be carried out ideally every year during the intercensital period. To facilitate longitudinal analyses from one year to another over the survey cycle, FAO sampling strategy for agricultural surveys recommends using either a panel or a partial sample rotation design. The panel design allows both cross sectional and longitudinal analyses with, in theory, all sample units. It is less costly and presents some operational advantages, as the enumerators shall interview the same holdings every year. However, the panel sample could suffer from attrition and obsolescence that would deter its representativeness and increase sampling errors. The partial rotation scheme is a great alternative to address the issue of sample attrition through a renewal of a part of the sample while allowing longitudinal analyses over two different survey occasions.

In case of a single-stage sampling, rotating samples are selected either directly in the population or in each stratum if a stratification is performed. In the framework of a multistage sampling, rotation is advised in the final selection phase. Accordingly, with a two-stage sampling, it would be recommended to rotate the secondary sampling units (SSU) rather than the primary sampling units (PSU). Graham (1963) recognises cost advantages associated with maintaining a fixed set of PSU although higher variability between them could be noticed in some cases and recommends definitively a rotation of higher-stage sampling units. In fact, rotating the PSU would be more expensive as it would imply updating more populations (populations of SSU in more

PSU and the population of PSU in each survey occasion). In addition, rotating SSU is likely to produce smoother estimates than rotating PSU.

The partial rotation design requires the selection of samples with partial overlaps that would correspond to the proportion of the sample planned to be renewed between to successive surveys. This can be performed using various techniques including the permanent random numbers (PRN) approach (Chromy 1979, Fan et al. 1962, Ohlsson 1992), the repeated collocated sampling or the rotation group sampling (Srinath and Carpenter 1995).

Fig. Example of a 10 years survey plan with a partial rotation design



2 Integrated agricultural and households surveys

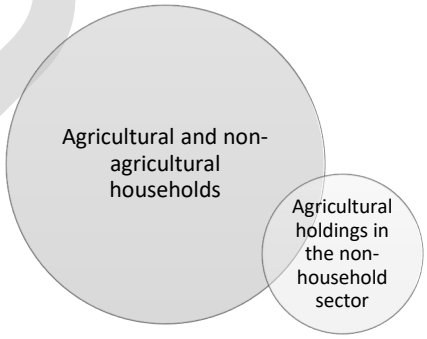
The integrated agricultural and households' surveys is a multipurpose survey with two main estimations goals: producing estimates on the whole population (agricultural and non-agricultural) of households (income, poverty etc.) and estimates of agricultural aggregates (planted area, production etc.) from all agricultural holdings (in both household and non-household sector).

2.1 Sampling frame

The ideal sampling frame would be a multiple frame of two lists: (i) complete list of households (agricultural and non-agricultural) and (ii) complete list of agricultural holdings in the non-household sector in the country.

The first list, that can be developed from a recent population and housing census, should include an information on the type of household: agricultural (denoted as A from now on) and non-agricultural households (denoted as B). The second list can be developed as described in section 1.2.

Fig. Sampling frame for the integrated agricultural and households' survey



2.2 Sampling design and sample size

A stratified simple random (or systematic) sampling without replacement is usually suitable for the population of agricultural holdings in the non-household sector. Usually it appears adequate to calculate the sample size based on the requirement of an accurate estimation of a proportion because of diversity and specialisation (livestock by species, crops, mixed…) of that population.

Regarding the population of households, a stratified two-stage sampling is cost effective in the context of most developing countries. The primary sampling units (PSU) are usually the enumeration areas used in a previous population census and the secondary sampling units (SSU) are households.

2.2.1 Size of SSU

The households' sample size should ensure a reliable estimation of key household related variables (e.g. income) in the population of households (A and B) and reliable estimation of agricultural related variables (e.g. planted area) from the sub population of agricultural households (A) as households in the sub population B do not practice agriculture.

To calculate the minimum sample size of households to fulfil this goal, the usual approximate formula based on the coefficient of variation can be used.

Let's consider for each estimation domain U_d:

- M_{Ad} and M_{Bd} total number of households respectively of type A and B.
- cv^2_{Aincd} and cv^2_{Bincd} coefficient of variation of respectively income of households of type A and B
- cv^2_{Aland} coefficient of variation of agricultural area of the agricultural household.
- cv^{*2}_d maximum relative error accepted for estimating the total (average) of income and agricultural area.
- $\widetilde{deff}_{Aincd}$, $\widetilde{deff}_{Bincd}$ and $\widetilde{deff}_{land}$ estimates of the design effect for respectively income of households of type A and B and agricultural area.
- g is the expected response rate.

The minimum sample size of households (m_d) in the domain U_d is:

$$m_d = \frac{1}{g}\left[Max\left(\widetilde{deff}_{land}\frac{cv^2_{Aland}}{cv^{*2}_d + \frac{cv^2_{Aland}}{M_{Ad}}}, \widetilde{deff}_{Aincd}\frac{cv^2_{Aincd}}{cv^{*2}_d + \frac{cv^2_{Aincd}}{M_{Ad}}}\right) + \widetilde{deff}_{Bincd}\frac{cv^2_{Bincd}}{cv^{*2}_d + \frac{cv^2_{Bincd}}{M_{Bd}}}\right]$$

Or

$$m_d = \max\left(m_{dA,inc}, m_{dA,land}\right) + m_{dB,inc} = m_{dA} + m_{dB,inc}$$

This procedure requires having all the variables in the formula for A and B type households (Ag Households and Non-Ag Households) in each domain *d*. However, it may happen for instance that the coefficient of variation of the income cannot be estimated for each sub population if the exercise is done with data from a household survey that did not cover agricultural activities. In such case if $m_{d,inc}$ is the overall minimum size of households for a reliable estimate of the income, we have:

$$m_{d,inc} = \frac{1}{g}\widetilde{deff}_{incd}\frac{cv^2_{incd}}{cv^{*2}_d + \frac{cv^2_{incd}}{M_{Ad} + M_{Bd}}}$$

And

$$m_d = \max\left(\widetilde{W}_{Ad} m_{d,inc}, m_{d,land}\right) + (1-\widetilde{W}_{Ad}) m_{d,inc}$$

Where:

- cv^2_{incd} is the coefficient of variation of the income of households in the domain *d*
- $\widetilde{deff}_{incd}$ is an estimate of the design effect for the income of households
- $\widetilde{W}_{Ad}$ is an estimate of the proportion of agricultural households in the domain *d*.

2.2.2 Size of PSU

When PSUs are selected with probability proportional to their size, selecting a fixed number of m_0 Households per PSU will allow having constant weights. This means that the number of PSUs to be selected in each estimation domain *d* would be given by dividing the sample size of households by m_0. With this approach, the number of PSUs to be selected in the domain *d* is given by:

$$n_d = \left[\frac{m_d}{m_0}\right] + 1$$

Where [ ] is the integer part.

The value of m_0, i.e. the fixed number of households to select in each PSU, is usually determined considering maximum enumerators' workload during survey implementation. An arbitrary value, generally varying between 10 and 15 is usually considered. Alternatively, it can be determined considering both costs and homogeneity of Households in the PSUs (intraclass correlation $\bar{\rho}$) (Kish (1965), equation 8.3.7):

$$m_0{}^* = \sqrt{\frac{c_p \times (1-\bar{\rho})}{c \times \bar{\rho}}}$$

where c_p and c are respectively the cost of adding an additional PSU into the sample and the unit cost of an interview. The intraclass correlation $\bar{\rho}$ can be estimated from previous surveys; since two variables are considered, consumption and land, the minimum value, $\bar{\rho} = min(\bar{\rho}_{cons}, \bar{\rho}_{land})$, should be considered (it is a conservative choice). It is worth noting that this formula is an approximation based on two stage simple random sampling of both PSUs and SSUs, when PSUs size does not vary greatly.

2.1 Stratification and allocation of PSUs

When implementing a two-stage sampling, FAO (2017) recommends a stratification of the EAs by administrative zones (e.g. regions, provinces, etc.) and agro-ecological zones before the first stage selection in order to improve the estimates of agricultural statistics. Stratification of PSUs should be carefully controlled since having too many strata is not desirable (risk of strata with too low sizes and cases of allocation of one unit per stratum that complicate variance calculation; see also Cochran 1977, page 132-134). To avoid too many strata, explicit stratification can be coupled with an implicit stratification.

When the list of households from the PHC is outdated, the actual structure of the households within the sampled PSUs can be known only after a fresh listing of households in these PSUs. A major drawback is the lack of control of the final sample especially the number of agricultural households required in the domain, since the selection is done at level of PSUs that may show a varying situation in terms of proportion of agricultural households.

In the context of the integrated agricultural and rural survey, to maintain control on the final sample size by household type (A and B) it would be preferable to make a first level stratification of the EAs in term of proportion of agricultural households in each of them estimated from the latest PHC or other suitable sources. Even if the PHC data are considered outdated, this structural information (proportion of agricultural households) may not likely vary too much in all PSU and could be helpful for stratification purposes. The first level stratification below may be considered using a proportion threshold ρ ($\frac{1}{2} < \rho < 1$).

First level PSU strata	**Definition**
Agricultural	$Proportion\ of\ agriculture\ households\ in\ the\ PSU\ \geq\ \rho$
Mixed	$1-\rho <\ Proportion\ of\ agriculture\ households\ in\ the\ PSU\ <\rho$
Non-agricultural	$Proportion\ of\ agricultural\ households\ in\ the\ PSU\ \leq\ 1-\rho$

The sample of PSUs in the domain d (n_d) can be allocated using parameters θ_a, θ_m and θ_{na} with $\theta_a + \theta_m + \theta_{na} = 1$

First level allocation	
First level PSU strata	**Allocation of the sample of PSU**
Agricultural	$\theta_a n_d$
Mixed	$\theta_m n_d$
Non-agricultural	$\theta_{na} n_d$

If m_0 households will be selected in each sampled PSU using an SRSWOR, the expected number of agricultural households in the final sample (m_{dAexp}) is:

$$m_{dAexp} = \rho m_0 \theta_a n_d + (1-\rho) m_0 \theta_m n_d + \delta m_0 \theta_{na} n_d = (\rho\theta_a + (1-\rho)\theta_m) m_0 n_d + \delta\theta_{na} m_0 n_d$$

$\delta < 1$ is unknown before the selection of the sample of households contrary to the other parameters that are fixed by the sample designer.

Let's consider τ the proportion of agricultural households in the planned sample.

$$\tau = \frac{m_{dA}}{m_d} = \frac{m_{dA}}{m_0 n_d} \Longrightarrow m_{dA} = \tau m_0 n_d$$

To ensure the achievement of the planned sample of agricultural households in the final sample of households, parameters θ_a, θ_m and θ_{na} could be fixed to have $m_{dAexp} \geq m_{dA}$. That corresponds to:

$$(\rho\theta_a + (1-\rho)\theta_m) m_0 n_d + \delta\theta_{na} m_0 n_d \geq \tau m_0 n_d$$

δ being unknown, parameters θ_a, θ_m and θ_{na} can be therefore fixed under the following conditions:

$$\rho\theta_a + (1-\rho)\theta_m \geq \tau$$

$$\theta_{na} = 1 - (\theta_a + \theta_m)$$

A second level stratification of PSU may be performed inside the first level strata (e.g. by agro-ecological zones, land use classes, size categories based on population, agricultural area…). The allocation in these second level strata can follow different criteria. Typically, in household surveys an allocation proportional to the population in the strata it is considered. FAO (2017) recommends the optimal allocation of Neyman for agricultural surveys. Kish (1987, p. 228) suggests a compromise solution between equal and proportional allocation:

$$n_{dh} = n_d \times \sqrt{\left(W_{dh}^2 + \frac{1}{H_d^2}\right)}$$

Where H_d is the number of strata in the domain *d*, while W_{dh} is the relative size of stratum *h* in domain *d*, it can be the proportion of PSU in stratum *h* compared to the domain total, $W_{dh} = N_{dh}/N_d$, (relative size in terms of population). A multivariate stratification and allocation could also be explored if the frame contains relevant variables correlated with households' income or agricultural area (household size, livestock, agricultural production, etc.) at PSU level.

3 Use of indirect sampling

3.1 Overview

In the framework of agricultural surveys, it happens often that the sampling units are different from the target observations units. This corresponds to an indirect sampling that usually complicates the calculation of the sampling weights of the target units of interests. A solution to this issue is the Generalized Weight Share Method (GWSM) developed by Lavallée (2007) that FAO suggests in contexts of indirect sampling when the direct calculation of the sampling of agricultural holdings is cumbersome. The method can be briefly described as follows. Let's suppose we are interested in a population U^B but for some reasons we do not have a sampling frame on that population. We will suppose also that there is a population U^A whose units have some linkages with the units of U^B and there is a sampling frame for the population U^A. An indirect sampling would simply consist in selecting a sample s^A from the population U^A and identify the units s^B of the population U^B that are linked to the sample units s^A. Units s^B become then our indirect sample to be surveyed and we need to calculate their sampling weights for estimations.

Let's consider

- $L_{ij} = 1$ if the unit i of U^B is linked to the unit j of U^A and $L_{ij} = 0$ otherwise.
- $L_i = \sum_{j \in U^A} L_{ij}$ is then the total number of links of the unit i with the population U^A
- w_j: sampling weight of unit j of s^A.

The sampling weight w_i of the unit i of s^B using the GWSM is then:

$$w_i = \frac{\sum_{j \in s^A} w_j L_{ij}}{L_i}$$

Lavallée (2007) proved that sampling weights calculated through the GWSM provide unbiased estimations on population U^B using the Horvitz–Thompson estimator. For that two important conditions should be fulfilled (Falorsi et al. 2015):

- The sampling strategy for the selection of s^A is roughly unbiased.
- Every unit in the population U^B has at least one link with the units of the population U^A.

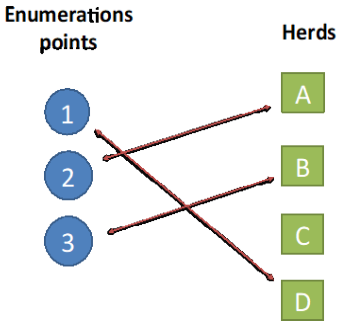
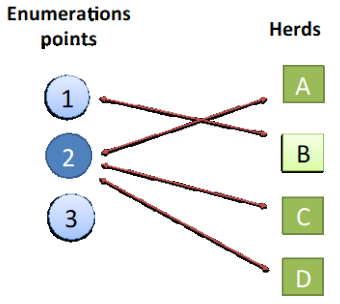
Consequently, it is important to define clearly the links and collect information on them during the survey for the use of the GWSM.

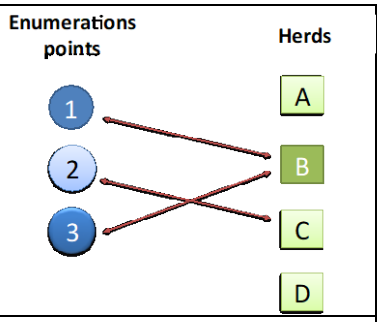
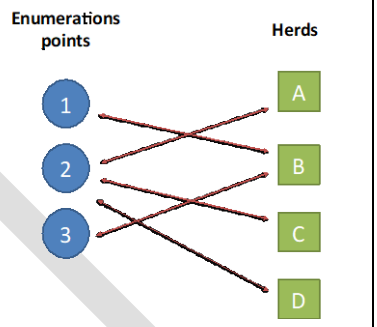
3.2 Case of mobile livestock survey

There are two types of mobile livestock: nomadic and semi nomadic. Nomadic livestock is by definition livestock raised by pastoralists not permanently settled and characterized by irregular, erratic movements that cover long distances. Semi-nomadic (transhumant) pastoralists are generally settled for a certain period of the year; their movements are regular, cyclical and span short distances. The livelihoods of both pastoralists depend almost entirely on livestock (FAO, 2016). In many developing countries, this type of livestock represents an important part of the total livestock. In 2016, FAO implemented methodological works on survey methodology for mobile livestock for improvement of livestock statistics in these countries.

There are two surveys approaches for the enumeration of nomadic and semi-nomadic livestock: aerial and ground surveys. Aerial surveys use existing sound methodology used in the counting of wild animals through low aircraft flights or taking special aerial photographs. In general, they are not suitable for livestock surveys in developing countries because of associated high cost and the focus on livestock numbers.

Ground surveys are more suitable to collect all the data needed for policy and interventions. However, there are many challenges to obtaining reliable information about transient livestock populations wandering across extensive rangelands in search of seasonally available pasture. For sampling surveys using the enumeration points (watering, vaccination, dipping points, livestock markets…), FAO recommends an indirect sampling approach. That would consist in first selecting a sample of enumeration points (watering points are preferred as all livestock frequent them) and then collect data on mobile herds at these points. This is a case of indirect sampling because of the possibility of complex linkage between enumeration points and mobile herds as presented below. The Generalized Weight Share Method (GWSM) can be used to calculate the sampling weights of herds using information on the watering points that they frequented during the year.

One-to-one Each enumeration point is linked to only one herd and vice versa i.e. each herd frequents one and only one enumeration point during the year. This kind of link is very rare in practice.	
One-to-many An enumeration point can be frequented by many herds but a herd can be linked to only one enumeration point. This kind of link is also rare in practice.	
Many-to-one Many enumeration points are frequented by one herd, but each herd can be linked to many enumeration point. This kind of link is more likely but quite rare in practice.	

Eventually, some rare exceptions may be faced with some specific enumeration points: for example, if in a country each nomadic herd generally uses the same stock route, then there will be a "many to one" type of link between herds and stock routes in that country.	
Many-to-many Many herds can be linked to one enumeration point and many enumeration points can be linked to one herd. Generally during a year, a transhumant or nomadic herd drinks in many watering points and a specific watering point is frequented by many herds. Thus, generally there is a "many to many" type of links between herds and enumeration points. This kind of link is certainly the most frequent in practice.	

3.3 Case of agricultural survey with an area frame

With an area, the sampling units are generally segments, points or cluster of points and then the agricultural holdings that have some land in the sampled segments or that have parcels of land in which the points are located are interviewed. This is another case of indirect sampling of agricultural holdings. In contexts where average number of distinct plots per agricultural holdings is important and located in different places, as usual in developing countries, multiplicities (multiple appearances) may occur in the indirect sample of holdings obtained through the area sample of points or segments. The GWSM described above can be used for the calculation of the holdings' sampling weights. The following definition of the link between area units and holdings may be adopted (FAO, 2017):

- A holding is linked to a segment if one of its parcels of land intersects the segment; and
- A holding is linked to a point if a portion (in terms of area or proprietary rights) or the total of the parcel in which the point is located belongs to it.

4 Subsampling and estimations

4.1 Overview

In agricultural and households surveys, subsampling can be used as a cost effective tool for various purposes including the following.

Use of different estimation domain for specific information

If the main domains of the survey are, as usual, sub-national administrative areas (regions, provinces, districts etc.), the country could consider that some information are just necessary for estimation at the national level. Therefore, it will not be necessary to collect them in the full sample in each estimation domain. The sample size for country level estimation can be calculated for these information and the corresponding questions will be administered only to a subsample in each estimation domain. For instance, if most information collected by

a rotating module are not requested for subnational policy making, the rotating questionnaires could be administered to a subsample.

Collecting information with high operation cost and/or time consuming

Some data collection methods provide high quality information but may their implementation may be too expensive in the full sample. Objective measurements of land (e.g. using GPS) or of yield (crop-cutting) may be considered as an example. Such operation may be perform on the subsample and the results can be used to correct measurement error for the whole sample.

4.2 Case of crop cutting

Crop-cutting is an operation particularly heavy that has significant effect on the survey time and budget. It requires the acquisition of specific equipment and at least two visits of the enumerator of the holding (during planting and harvesting periods). In some countries (e.g. Niger, Burkina Faso), the full sample of holdings is considered for crop-cutting and all plots of covered by the operation. However, a subsample of holdings and/or plots may be used especially when only national level estimates of crop yields are expected from the survey or if the results of crop-cutting are aimed to be used for correcting yield collected by farmers' declarations.

4.2.1 Subsampling approaches

Options for subsampling for crop-cutting include:

- Selecting directly a subsample of plots for implementing the crop-cutting. This is an efficient option (in term of quality of estimations) but has some operational constraints. In fact, it can be performed only after the listing of all plots of the holdings. Therefore this listing should be completed timely to allow the processing of the data and selection of the subsample of plots. In addition, large plots or plots very from the holding dwelling may appear in the sample increasing operation costs.
- Selecting a subsample of holdings and covering all or some plots of each subsampled holding by the crop-cutting operation. This option may allowing covering much more parcels than the previous at a lower cost as the holding is a cluster of plots. However, it is less efficient than the previous as cluster sampling leads to higher variance.

4.2.2 Subsample size

The size of the subsample of holdings/plots can be calculated under budget constraint. For instance, in Senegal, a subsample of 60 plots are selected in each district for implementing the crop-cutting. However this approach does not guarantee a reliable estimate of the yield.

The size of the subsample can be also calculated using the variability of the yield of a key crop. If the cv_y is the coefficient of variation of the yield, estimated from a previous survey and cv^* the maximum relative error expected, the following formula can be used.

$$size = \frac{cv_y^2}{cv^{*2}}$$

4.3 Case of farm level post-harvest losses

The FAO's Guidelines on the measurement of harvest and post-harvest losses discusses the main postproductions operations during which harvest losses occur at the different stages of the value. At the farm level, in the case of grains (cereals and pulses) losses occur mainly during the following operations: harvesting, threshing or shelling, cleaning or winnowing, drying and storage in the holding (FAO, 2018a). The Guidelines recommend using probability sample surveys as the backbone of any loss assessment, complemented by other methods that may be used mainly as preliminary assessments or to further analyze certain aspects related to PHL. With such surveys, loss measurements can be (i) objective – drawn from crop-cutting on the field or laboratory analysis of grain sampled from storage facilities – or (ii) subjective, by asking the respondent (farmer, storage facility manager, etc.) to provide his or her own estimate of loss.

In the framework of the initiative, it is recommended to make post-harvest losses assessments using a subsample for a number reasons including reducing operation costs and respondent burden. In fact, both options of loss measurements (objective and subjective) are relatively expensive and time-consuming, and require well-trained personnel (FAO, 2018a). Objective measurements are particularly expensive and require many visits of the enumerators to the farm. Subjective assessments are cheaper but require additional visits to the farm in particular for collecting information on storage losses; thus, additional cost may be important. In addition, field tests performed by FAO in the framework of the Global Strategy showed important measurement errors when comparing objective and subjective measurements in Ghana, Malawi, Namibia and Zimbabwe (FAO, 2018b).

Another reason for subsampling is that this information is generally expected at national level. Subnational domain estimations are therefore not particularly required. In addition, regarding international demand, SDG 12.3.1 on Global Food Loss Index is expected at country level.

Important indicators related to post-harvest losses are the proportions of losses at crop and operation level. The crop of interest should be specified, they are generally cereals and pulses. Therefore, the subsampling should be performed among holdings producing the target crops. Proportions being of interest, the minimum size of the subsample for reliable estimate at the national level can be calculated using a formula based on the estimation of a proportion as proposed below.

$$n_{subsample} = \mathcal{Z}^2_{\alpha/2} \frac{\hat{p}(1-\hat{p})}{\varepsilon^2}$$

Where

- $\hat{p}$ is the value or estimation of a proportion of losses of a key crop (a value of 0.5 gives the maximum size)
- $\mathcal{Z}_{\alpha/2}$ is the z score for (1−α)100% confidence interval
- ε is the maximal absolute error accepted

4.3.1 Estimations with subsampling

If subsampling is used to collect a specific information, estimations can be done using (i) a three-stage sampling scheme or (ii) a two-phase sampling perspective.

If all the sample of holdings are covered and a subsample of plots selected in each of them, then that corresponds to a three-stage sample selection. New sampling weights should be calculated for the subsampled

plots by multiplying the inverse of their probability of selection and the sampling weights of holdings. Estimators of variances provided for the two-stage design can still be used here with few adaptations.

However, selecting a subsample of holdings would correspond to a two-phase sampling. Estimations can be done using regression or ratio estimators. Regression estimators are considered more efficient in this context (Cochran, 1977). Let's consider the use of subsampling for objective measurements (GPS measurement or crop cutting) in a two phase scheme (subsample of holdings). For simplicity we will consider the case of a sample holdings selected through a simple random without replacement from which a simple random subsample of holdings is selected for the objective measurements. Let's consider y the yield measured using the crop-cutting in the subsample and x the yield collected by declaration on the whole sample. From Sitter (1997), the regression estimator used to estimate a more accurate average yield is:

$$\bar{y}_{reg} = \bar{y} + \beta(\bar{x} - \bar{x}_{subsample})$$

Where: $\bar{x}_{subsample}$ is the average of x in the subsample and β the least squares regression coefficient of y on x in the subsample.

An estimator of the variance of $\bar{y}_{reg}$ is

$$\tilde{v}(\bar{y}_{reg}) = \left(\frac{1}{n_{subsample}} - \frac{1}{n}\right) s_d^2 + (\frac{1}{n} - \frac{1}{N}) s_y^2$$

Where:

- $n_{subsample}$: size of the subsample
- n : size of the whole sample
- N : size of the population of holdings
- s_d^2 is the sample variance of the quantities $d_i = y_i - \bar{y}_i - \beta(x_i - \bar{x}_i)$

References

Chromy, J. R. (1979). *Sequential sample selection methods*. Proceedings of the American Statistical Association Section on Survey Research Methods of the American Statistical Association, 401- 406

Cochran, W.G. (1977). *Sampling Techniques*. 3rd Edition. John Wiley & Sons: New York, USA.

Falorsi, P.D. Bako, D. Righi, P. Piersante, A. (2015). *Integrated Survey Framework*. FAO Publication. Rome

Fan, C. T., Muller, M. E. Rezucha, I. (1962). *Development of Sampling Plans by Using Sequential (Item by Item) Selection Techniques and Digital Computers*. Journal of the American Statistical Association, 57, 387-402.

FAO (2015a). *World Census of Agriculture 2020. Volume 1: Programme, concepts and definitions*. FAO Publication. Rome.

FAO (2015b). *Handbook on Master Sampling Frames for Agricultural Statistics*. FAO Publication. Rome.

FAO (2016). *Guidelines for the Enumeration of Nomadic and Semi-Nomadic (Transhumant) Livestock*. Rome. August 2016

FAO (2017). *Handbook on the Agricultural Integrated Survey (AGRIS)*. GSARS. Rome

FAO (2018a). *Guidelines on the measurement of harvest and post-harvest losses. Recommendations on the design of a harvest and post-harvest loss statistics system for food grains (cereals and pulses).* FAO Publication. Rome

FAO (2018b). *Accelerated Technical Assistance Plan for Africa. Global Office Final report*. Technical Report Series GO-44-2018. FAO Publication. Rome. Available at http://gsars.org/en/accelerated-technical-assistance-plan-for-africa-global-office-final-report/

Graham, J. E. (1963). *Rotation designs for sampling on successive occasions.* Retrospective Theses and Dissertations. Paper 2384.

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

Kish, L. (1987). *Statistical design for research*. New York, NY: John Wiley & Sons.

Ohlsson, E. (1992). *SAMU, The system for Co-ordination of Samples from the Business Register at Statistics Sweden-A methodological description*, R&D Report 1992: 18, Stockholm: Statistics Sweden

R. R. Sitter (1997). *Variance Estimation for the Regression Estimator in Two-Phase Sampling*, Journal of the American Statistical Association, 92:438, 780-787

Srinath, K.P. Carpenter, R.M. (1995). *Sampling methods for repeated business surveys*. In Business Survey Methods, edited by Brenda Cox et al., pp 171-183. Wiley, New York.