Food and Agriculture Organization
of the United Nations

MASTER SAMPLING FRAME(MSF)

FOR

AGRICULTURAL STATISTICS

Module 3 – Session 4:

Survey design and Estimation when the MSF is a Multiple Frame

2017

Objectives of the presentation

- Introduce the concept of multiple frame sampling
- Discuss the principles of multiple frame sampling
- Deal with issues arising from the use of multiple frame sampling
- Discuss the estimation in the context of multiple frame

Outline

- Principles of multiple frame sampling
- Problems in the application of multiple frame surveys
- Dual frame estimator
  - Hartley and the screening estimator
  - The Fuller-Burmeister estimator
  - The Skinner-Rao estimator
  - Single frame-type estimator
  - Choosing among dual-frame estimators
- Estimation of domain parameters
- Using auxiliary information
- Allocation of sample size to frames

Introduction

- Rational: problem of coverage (missing in-scope units):
  - Difficulty to provide with the list of the entire units of the (target) population with a single frame.
  - In other word: none of the single frames could cover the entire target population.
- One of the solution is the Multiple Frame:
  - In some situations, coverage can be improved by using more than one sampling frame
  - Indeed several frames together have better coverage than a single frame
  - The gain is that one can exploit the strengths and offset the weaknesses of each type of frame.

WHAT IS AN MULTIPLE FRAME?

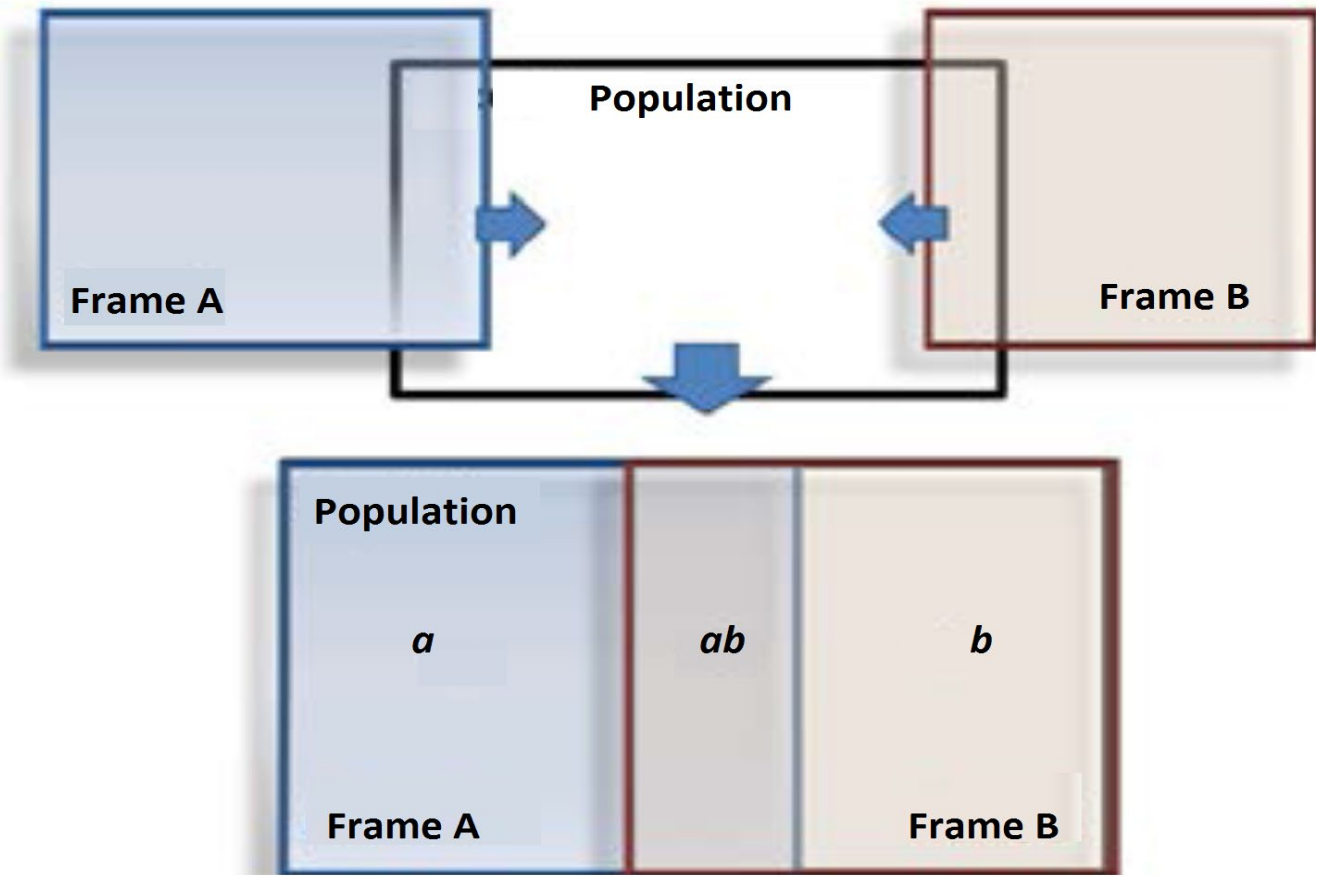
1.1 Definition of a Multiple Frame?

- Multiple frame sampling involves the joint use of two or more sample frames.
- For agricultural purposes, this usually involves the joint use of area and one or more list frames.
- Example: Agricultural sector:
  - Use of several lists frames
    - List of food crop farmers, list of cash crop farmers…
  - Use of area and list frames

1.2 Examples of Multiple Frame:

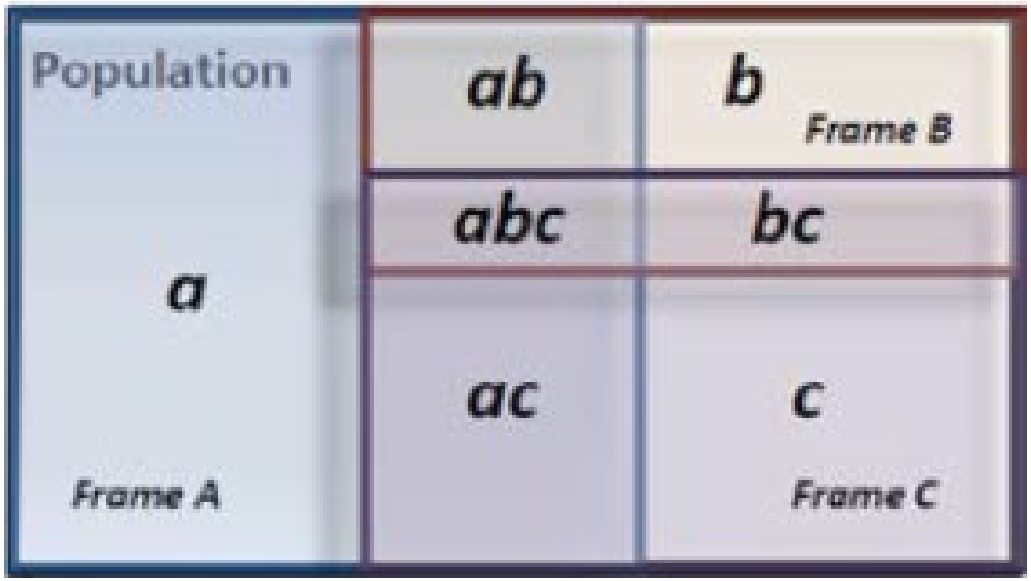
- Iachan and Dennis (1993): sampling of the homeless population in Washington DC (See also Lohr 1999 p.402).
  - Three frames were considered:
    - Homeless shelters
    - Soup kitchens and
    - Encampments and streets

1.2 example of schema: dual frame



1.2 example of schema

- Three frames



1.3 Two main requirements for the use of Multiple Frame

- There are two important requirements:
  - Completeness, and
  - Identifiability

- The **completeness**: the union of both frames provide full coverage for the target population. In this way, every element is listed in at least one of the frames. *(refer to next slide)*

- The **identifiability**: for any sampled element, it is possible to understand whether or not it belongs to one of the frames.

1.3 Two main requirements for the use of Multiple Frame (cont'd)

- **Completeness:**
  - Every farm in the population belongs to at least one frame
  - The concept of "completeness" involves two aspects: coverage and information provided for each frame unit
  - The area frame is used to ensure the completeness of the master frame because it is capable of covering all farms and their land

11

1.3 Two main requirements for the use of Multiple Frames (cont'd)

- **Identifiability:**
  - For any sample unit from any frame, it is possible to determine whether the reporting unit belongs to any other frame
  - The requirement of identifiability is met by determining which area frame reporting units can also be selected from the list frame

12

Advantages and problems in the application of Multiple Frame Survey

2.1 Advantages of Multiple Frame

- Build on strengths of each frame and minimize their weaknesses.
- In case of Area frame combine with a list frame:
  - MF Allows the easy and not expensive creation of lists of agricultural holdings only in the selected areas, instead of making it in the entire country.
  - Data collection can be inexpensive because sample units are conglomerated in the selected areas, instead of being spread in all the country´s territory.
- Variability can be controlled and measured.
- Enable the study of special or rare products.

2.2 Problems in the application of multiple frame survey

- **Multiple frame surveys include all the complexities of single frame surveys**
  - All farms in the list frame must be completely identified by name, address, and any other name forms that can identify the farming unit.
  - The need to match names from the area frame sample to names on the list frame complicates the survey process and is subject to non-sampling errors.
  - Mapping the area frame sampling unit onto a reporting unit, as does the list frame name. It is essential that a name be associated with the area tract.
  - Choice between many available lists. For example, one list of names may derive from the agricultural census and another from an administrative source. Choice one list or combining them?

2.2 Problems in the application of multiple frame survey

- **Some practical tips**
  - **The need to identify all domains when there are two or more list frames** greatly complicates the survey and estimation process. For this reason, it is more practical to combine all lists and remove duplicates prior to sampling.
  - **A common problem is the temptation to make the list as large** as possible to avoid the occurrence of outliers. However, the larger the list, the more subject it is to duplication. The smaller the list, the easier it is to avoid duplication and to determine the non-overlap domain.
  - **Another common problem is the temptation to add names found in the area** frame survey to the list frame. These additions introduce a downward bias, because the estimation probabilities have been changed (reduced) when added to the list.

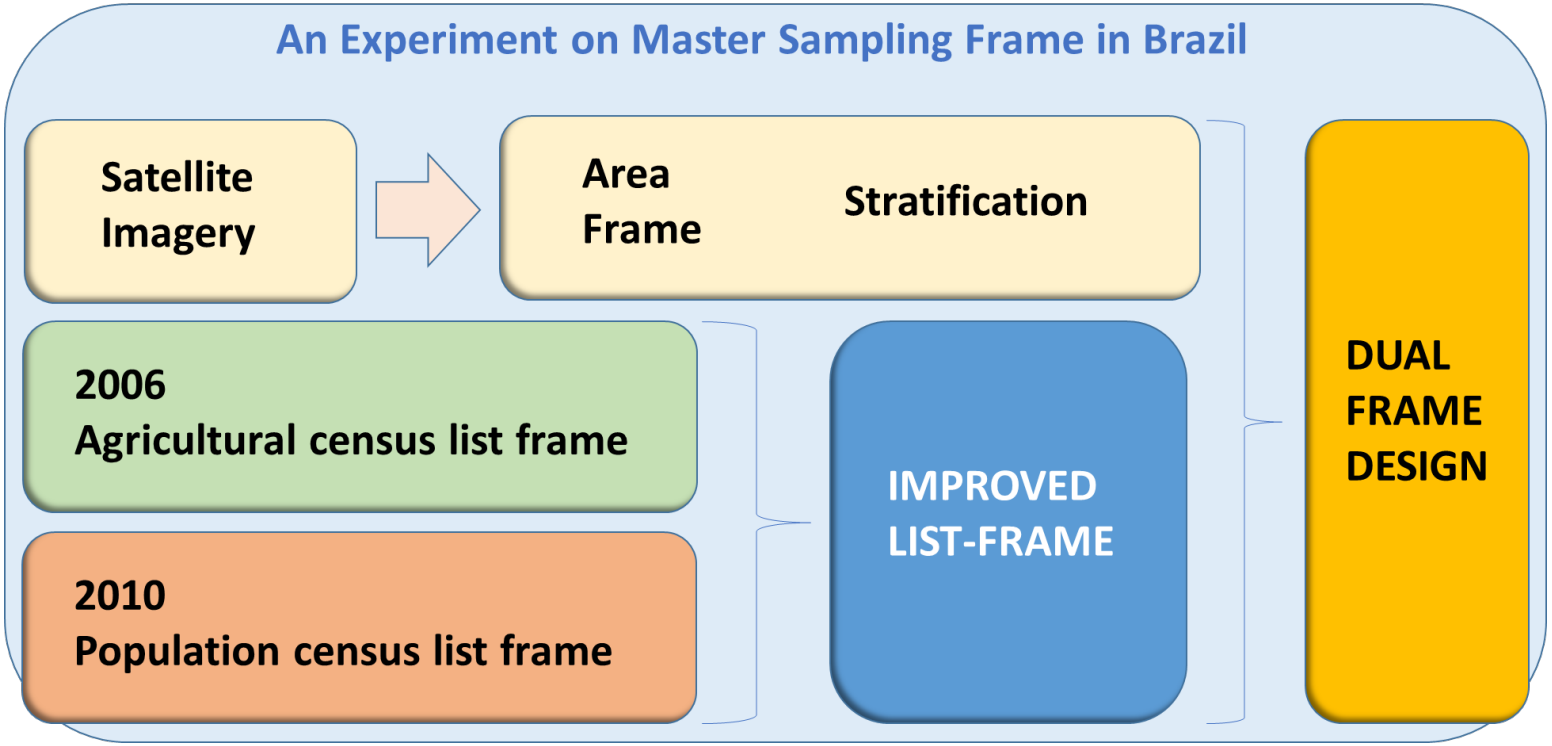
2.2 Problems in the application of multiple frame survey: Some practical tips

- Identify two domains for the area frame:
  - 1st Domain: those that are not on the list (non-overlap)
  - 2nd Domain: those that are on the list (overlap).
- Assumption of identification:
  - Every farm or household included in the LF has a chance of being selected in the AF sample.
  - Identification of domains must be based on the area frame sampled units, and not on the entire area frame.
  - If a name from the AF does not match a LF name exactly, but is close, then it may be possible to compare addresses or secondary names associated with the reporting units.

2.3 Multiple frame sampling: Brazil

- The MSF is being used for the System of Integrated Household Surveys, in which all individual household surveys use the same frame.
- The area sample frame was constituted by strata of the land use established according to the rate of cultivated land, or by the predominance of crops.
- Based on the 1985 Agricultural Census information, several lists of holdings that concentrate a large percentage of the total of the variable were constructed.
- These lists, including a relatively small number of holdings, are called Lists of Special Holdings and are updated every year.
- For a given variable, the multiple frame estimator is the sum of the estimators from both samples, the area sampling estimator and the list sample estimator based on the list frame of special holdings.

2.3 Multiple frame sampling: Brazil



2.4 Summary

- This module provides the principles of multiple frame sampling.
- It gives an overview of common problems in the application of multiple frame sampling due to the requirement of identifying the overlap between frames.
- While the term “multiple frame sampling” implies that more than two frames can be used, the complexities in determining the overlap between frames appears to favour limiting the choice to two frames.
- The general conclusion is that the list frame containing mostly large commercial farms and farms producing important but rare items should be kept as small as possible

Dual frame estimators

Main dual frame estimators

- Hartley and the screening estimator
- The Fuller-Burmeister estimator
- The Skinner-Rao estimator
- Single frame-type estimator

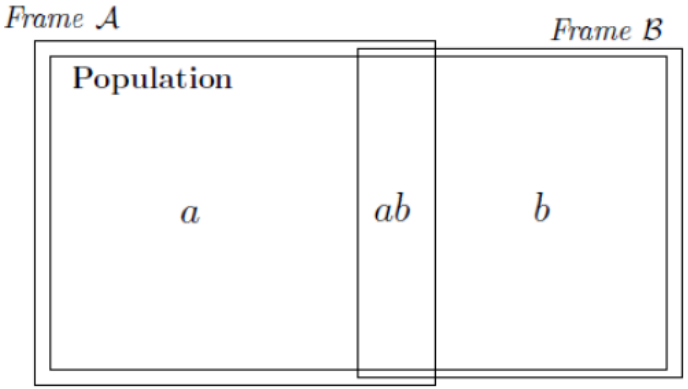
3.1 Dual frame estimator: background

- Hartley (1962) was the first to theoretically derive the multiple frame methodology.
- Considerable extensions to this work are due to Cochran (1965), Lund (1968), Fuller and Burmeister (1972), Hartley (1974), and Bosecker and Ford (1976).
- Example of the use of multiple frame: the sampling of the homeless population in Washington DC (Iachan and Dennis 1993, See also Lohr 1999 p.402).
  - Three frames were considered:
    - Homeless shelters
    - Soup kitchens
    - Encampments and street

3.1 Dual frame estimator: background (cont'd)

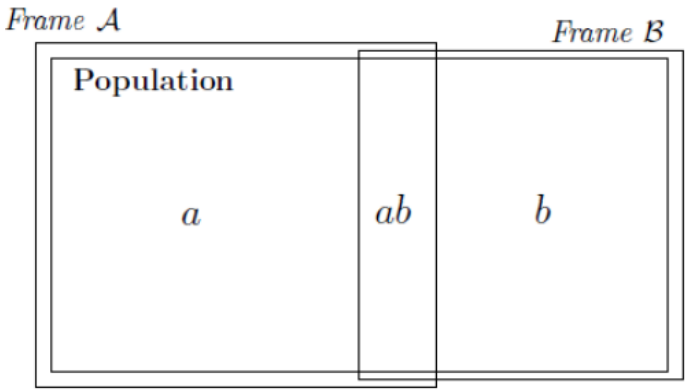
- The basic theory of multiple frame sampling (Hartley, 1962; Kott and Vogel, 1995) begins with dividing the population into mutually exclusive domains.

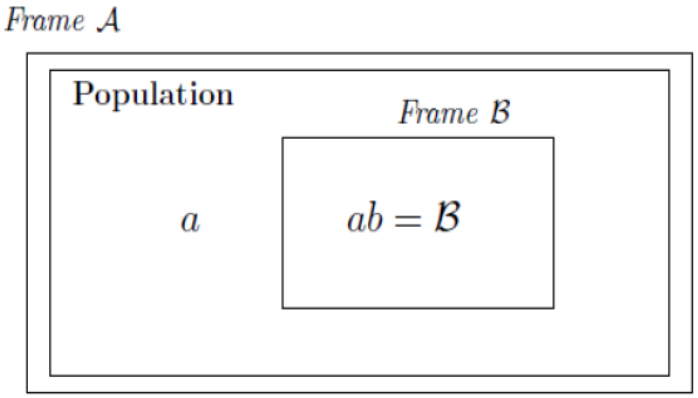
- inference in dual frame surveys considers the broad scenario illustrated by the figure below, in which three domains can be identified: *a b* and *ab.*



3.1 Dual frame estimator: background (cont'd)

- The dual frame method combines (for example) independent samples from an incomplete list frame and an area frame that is assumed to be complete.
- both situation below could be considered





3.1 Dual frame estimator: background (cont'd)

- Let consider the formula below:
  - π_k^A and π_k^B be the first−order inclusion probabilities for the elements of each frame in a dual frame survey
  - y_k be the value of the variable of interest for $k \in U$
  - U_{ab} denote the set of population elements belonging to domain *ab*
  - While S_{ab}^A denotes the sample set of elements from U_{ab} selected from frame A
  - $N_A=N_a+N_{ab}$ *and* $N_A=N_a+N_{ab}$

3.2 Dual Frame Estimators: Hartley's estimator

- Hartley (1962) : can be expressed as a weighted average between the appropriate Horvitz-Thompson (Horvitz and Thompson 1952) estimators applied to each dual frame domain:

- $\hat{t}_H = \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \sum_{k \in S_b} \frac{y_k}{\pi_k^B} + \sum_{k \in S_{ab}} y_k^*$

- $y_k^* = \begin{cases} p \frac{y_k}{\pi_k^A}, if \quad k \in S_{ab}^A \\ (1-p) \frac{y_k}{\pi_k^B}, if \quad k \in S_{ab}^B \end{cases}$

- and p is a weighting constant ($0 < p < 1$) chosen to minimize the variance of the estimator $\hat{t}_H$

3.2 Dual Frame Estimators: the screening estimator

- Considering an agricultural survey with Frame A (area frame) and Frame B (list frame), and B is included in A ($B = A \cap B$), Hartley's screening estimator is:

- $\hat{Y} = \hat{Y}_a + \hat{Y}_{ab} = \hat{Y}_a + \hat{Y}_L$, (general expression)

- $\hat{t}_H = \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \sum_{k \in S_{ab}} y_k^*$

- $\hat{t}_H = \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \sum_{k \in S_b} \frac{y_k}{\pi_k^B}$

- $\hat{t}_H = \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + p \sum_{k \in S_{ab}} y_k^* + (1 - \mathrm{p}) \sum_{k \in S_b} \frac{y_k}{\pi_k^B}$

3.2 Dual Frame Estimators: the screening estimator

- Supposing that a simple random sampling is applied to each frame, and denoting the domain population variances by σ_a^2, σ_b^2 and σ_{ab}^2 is approximated by the variance for stratified samples with an allocation proportional to the domain sizes (ignoring the finite population correction factors)

$$Var(\hat{t}_H) = \frac{N_A^2}{n_A}\left[\sigma_a^2\left(1-\frac{N_{ab}}{N_A}\right)+\phi^2\sigma_{ab}^2\frac{N_{ab}}{N_A}\right]+\frac{N_B^2}{n_B}\left[\sigma_b^2\left(1-\frac{N_{ab}}{N_B}\right)+(1-\phi)^2\sigma_{ab}^2\frac{N_{ab}}{N_B}\right].$$

$$\hat{N}_{ab}^A = \sum_{k\in S_{ab}^A}\frac{1}{\pi_k^A}, \qquad \hat{N}_{ab}^B = \sum_{k\in S_{ab}^B}\frac{1}{\pi_k^B}$$

3.3 The Fuller- Burmeister Estimator

- Fuller and Burmeister (1972) proposed an estimator that uses sample information from the frames to estimate N_{ab}
- Then, the Fuller-Burmeister estimator is given by
- $\hat{t}_{FB} = \sum_{k\in S_a} \frac{y_k}{\pi_k^A} + \sum_{k\in S_b} \frac{y_k}{\pi_k^B} + \sum_{k\in S_{ab}} y_k^* + p_2\left(\widehat{N}_{ab}^A - \widehat{N}_{ab}^B\right)$
- $y_k^* = \begin{cases} p_1 \frac{y_k}{\pi_k^A}, if \quad k \in S_{ab}^A \\ (1-p_1)\frac{y_k}{\pi_k^B}, if \quad k \in S_{ab}^B \end{cases}$
- p_1 *and* p_2 are weighting constant ($0< p_1, p_2 < 1$) chosen to minimize the variance of the estimator $\hat{t}_{FB}$

$\widehat{N}_{ab}^A = \sum_{k\in S_{ab}^A} \frac{y_k}{\pi_k^A}$ and $\widehat{N}_{ab}^B = \sum_{k\in S_{ab}^B} \frac{y_k}{\pi_k^B}$

3.4 Single Frame Type Estimator

- Authors: Bankier (1986) and Kalton and Anderson (1986)
- This estimator relies on a set of sampling weights which enable the estimator to be written as the sum of only two Horvitz-Thompson estimators, each covering the sample data from one of the frames. It is written as shown below:
- $\hat{t}_B = \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \sum_{k \in S_b} \frac{y_k}{\pi_k^B} + \sum_{k \in S_{ab}} \frac{y_k}{\pi_k^A + \pi_k^B}$

3.5 The Skinner- Rao Estimator

- Authors: Skinner and Rao (1996)
- It is a pseudo-maximum likelihood estimator (PML) that uses a single set of weights in each frame. This estimator can be expressed as
- $\hat{t}_{FB} = \frac{N_A - \hat{N}_{ab,SR}}{N} \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \frac{N_B - \hat{N}_{ab,SR}}{N} \sum_{k \in S_b} \frac{y_k}{\pi_k^B} + \sum_{k \in S_{ab}} \gamma y_k^*$
- $y_k^* = \begin{cases} \frac{y_k}{\pi_k^A}, if \quad k \in S_{ab}^A \\ \frac{y_k}{\pi_k^B}, if \quad k \in S_{ab}^B \end{cases}, , \gamma = \frac{\hat{N}_{ab,SR}}{\pi_k^A \hat{N}_{ab}^A + \pi_k^B \hat{N}_{ab}^B}$

3.6 Choosing among dual-frame estimators

- The estimators presented above display differing levels of complexity, depending on how they provide estimates for the overlapping domain *Yab*.
- It is recommended that the estimator be chosen on the basis of simplicity.
- Screening estimators are the simplest to understand and apply in practice, leaving the more complex aspects to the data collection and matching process.

3.6 Choosing among dual-frame estimators

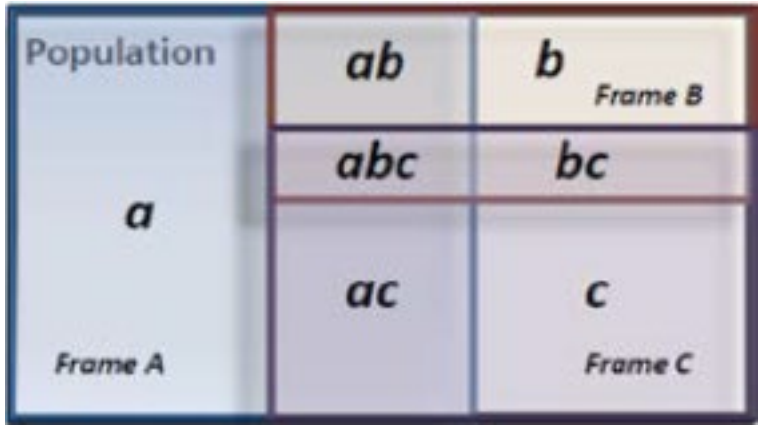
- The precision of estimates can be improved by investigating the feasibility and efficiency of other estimators based on simulation studies at a later stage.
- These studies seeking to compare the statistical performances of dual-frame estimators should take into account the peculiarities of each country and specific probability sample designs.
- To improve precision, one should not only search for competitive dual-frame estimators, but also ascertain how to incorporate auxiliary variables – that may be available through at least one of the frames – into the inference process.

Estimation of domain parameters

4. Estimation of domain parameters

- For example if three overlapping frames are used simultaneously in a multiple frame approach, seven (2^{3-1}) domains would be created,
- In general, for n overlapping frame, it would be 2^{n-1} domains

- Domains would be identified, and making inferences for each of these is strategic in producing estimates according to a multiple frame sampling approach.



4. Estimation of domain parameters (cont'd)

- Let Domain *a* be chosen to illustrate the estimators.
- Assume that a probability sample was selected from Frame A, are as follow:
- $\pi_k^A = P(k \epsilon S_A)$, the probability that unit *k* is included in the sample (first-order inclusion probability)
- $\pi_{kl}^A = P(k, l\ \epsilon S_A)$, the probability that units *k* and *l* are both in the same sample (second-order inclusion probability).

4.1 Horvitz-Thompson domain estimator

- If the domain size *Na* is unknown, then the following Horvitz-Thompson type estimator is recommended (Sarndal, Swensson and Wretmann, 1992; page 390):
- *For the domain total:* $\hat{Y}_a = \sum_{k \epsilon S_A} \frac{y_k}{\pi_k^A} \delta_{ak}$
- *For the domain mean:* $\bar{Y}_a = \frac{\hat{Y}_a}{\hat{N}_a}$
- Where $\delta_{ak} = \begin{cases} 1, if\ k \in a \\ 0, if\ not \end{cases}$ And $\hat{N}_a = \sum_{k \epsilon S_A} \frac{\delta_{ak}}{\pi_k^A}$
- It variance and variance estimator can be written as follow:
- $V(\hat{Y}_a) = \sum_{k \epsilon A} \sum_{l \in A} (\pi_{kl}^A - \pi_k^A \pi_l^A) \frac{y_k}{\pi_k^A} \frac{y_l}{\pi_l^A} \delta_{ak} \delta_{al}$
- $\widehat{V(\hat{Y}_a)} = \sum_{k \epsilon S_A} \sum_{l \epsilon S_A} (\pi_{kl}^A - \pi_k^A \pi_l^A) \frac{y_k}{\pi_k^A} \frac{y_l}{\pi_l^A} \frac{\delta_{ak} \delta_{al}}{\pi_{kl}^A}$

4.2 The π-weighted domain estimator for the total:

- On the other hand, if the domain size *Na* is known, then the π-weighted estimator is known to provide a better statistical performance (Sarndal, Swensson and Wretmann, 1992: p. 390): $\hat{Y}_a = N_a \bar{Y}_a$
- The variance of the π-weighted domain estimator can be approximated by:
- $V(\hat{Y}_a) = \sum_{k \epsilon A} \sum_{l \in A} (\pi_{kl}^A - \pi_k^A \pi_l^A) \frac{y_k - y_a}{\pi_k^A} \frac{y_l - y_a}{\pi_l^A} \delta_{ak} \delta_{al}$
- $\widehat{V(\hat{Y}_a)} = \sum_{k \epsilon S_A} \sum_{l \epsilon S_A} (\pi_{kl}^A - \pi_k^A \pi_l^A) \frac{y_k - \bar{Y}_a}{\pi_k^A} \frac{y_l - \bar{Y}_a}{\pi_l^A} \frac{\delta_{ak} \delta_{al}}{\pi_{kl}^A}$
- $\bar{Y}_a = \sum_{k \epsilon A} \frac{y_k}{N_a} \delta_{ak}$

Using auxiliary information

5.1 Using auxiliary information

- What is an auxiliary information or data?
  - Any kind of relevant, accurate and useful information that can be use to improve accuracy of decision or estimate
- Importance: useful to improve survey design for example:
  - Stratification
  - Sampling design
  - Allocation of sample size (e.g. per strata)
  - Guide sampling procedure (PPS or MPPS)
  - choice of method of selection

5.1 Using auxiliary information

- Source of auxiliary data
  - Other data sources
    - Population or agricultural census,
    - Any relevant survey data
  - Data from administrative reporting systems
    - Annual report
    - Activity report
    - Statistical Yearbook
- It very useful when a frame contain auxiliary information to improve the sampling designs and estimators

5.2 Examples of auxiliary informations

- Total agricultural population size per EA
- Total area under crop of a specific region
- Total production for a specific crop in a particular Region
- Standard deviation for a specific variable
- Variable's Average (age, production, area,…)

5.2 Use of auxiliary informations:

- Extracted from the Handbook on MFS,2015:
  - "Given the availability of (external) auxiliary information from at least one of the frames, it is desirable to take advantage of these either by incorporating this information into the sampling scheme, using PPS or MPPS designs, or by using it to assist in the construction of a more appropriate regression-type estimator."
  - "A list of names without any identifying information on the characteristics of the farm and its relative size in terms of land area and number of livestock will be no more statistically efficient than a list of segments or points from the area frame. The primary reason for adopting a list frame is to use its auxiliary information for sampling purposes."
  - "PSUs are selected with PPS, the auxiliary variable being cropping intensity (Delincé, 2015)."

5.2 Use of auxiliary information:

- Extracted from the **Technical Report Series GO-08-2015: Linking Area and List Frames in Agricultural Surveys**
  - "The choice of sampling scheme depends on the nature of the frame component units in use, as well as on the nature of any auxiliary information that may be available"
  - "Auxiliary information may be available in list frames, which would enable the use of efficient sampling schemes such as stratified sampling, probability proportional-to-size sampling or even both of these, as well as that of calibration and regression-type estimators."
  - "Identifying methods to improve the matching of several sources of data at the level of frame component units may support the feasibility of good quality frames, not only in terms of coverage rates but also of providing access to auxiliary information."

Allocation of sample size to frames

How to allocate sampling size between Area Frame and List Frame?

Allocation of sample size to frames:

- In dual-frame surveys, given an estimated sample size, the problem of sample size allocation to the frames must still be addressed.
- Suppose that the screening estimator is used and a simple random sample (without replacement) is taken in both Frames *A* and *L*. Then,
- $\hat{Y}_S = \hat{Y}_\alpha + \hat{Y}_L *$

Refer to slid about Horvitz-Thompson domain estimator

Allocation of sample size to frames

- Now, the problem of dual frame allocation is the same as the problem of determining the values for n_A and n_L that will minimize $Var(\hat{Y}_S)$ subject to cost constraints.

$$Var(\hat{Y}_S) = Var(\hat{Y}_a) + Var(\hat{Y}_L) = \left(\frac{N_a^2 \sigma_{ay}^{2*}}{n_A} - N_A \sigma_{ay}^{2*}\right) + \left(\frac{N_L^2 \sigma_{Ly}^2}{n_L} - N_L \sigma_{Ly}^2\right)$$

Lets assume the total cost *C* involved in the survey is such that

$$C = c_0 + n_A c_A + n_L c_L$$

where c_0 is a fixed overhead cost and c_A and *Cl* represent the cost of sampling and observing one element of Frames *A* and *L* respectively.

$\sigma_{\alpha y}^{2*} = P_a(\sigma_{\alpha y}^2 + Q_a \mu_{\alpha y}^{2+})$, $P_a = \frac{N_a}{N_A}$, Q_a=1-P_a

Allocation of sample size to frames

- In these conditions, it is known that the optimum allocation is to choose

$$n_A = n\frac{\sqrt{\frac{N_\alpha^2\sigma_{\alpha y}^{2*}}{c_A}}}{\sqrt{\frac{N_\alpha^2\sigma_{\alpha y}^{2+}}{c_A}}+\sqrt{\frac{N_L^2\sigma_{Ly}^2}{c_L}}}$$

$$n_L = n\frac{\sqrt{\frac{N_L^2\sigma_{Ly}^2}{c_L}}}{\sqrt{\frac{N_\alpha^2\sigma_{\alpha y}^{2+}}{c_A}}+\sqrt{\frac{N_L^2\sigma_{Ly}^2}{c_L}}}$$

$\sigma_{\alpha y}^{2*}=P_a(\sigma_{\alpha y}^2+Q_a\mu_{\alpha y}^{2+})$, $P_a=\frac{N_a}{N_A}$, Q_a=1-P_a

Summary

- Multiple Frame sampling involves the joint use of two or more sample frames.
- For agricultural purposes, this usually involves the joint use of area and one or more list frames.
- Two important requirements (Completeness and Identifiability) are necessary for the use of Multiple Frame

Summary (cont'd)

- The Multiple Frame take advantage of the strengths of each frame and minimize their weaknesses.

- Several estimates with different level complexity are developed in the context of Multiple Frame, but it is recommended that the estimator be chosen on the basis of simplicity. Screening estimators are the simplest to understand and apply in practice.

- In addition, when allocation sampling size to the area and list frame, the one to be defined should minimize the variance $Var(\hat{Y}_S)$ subject to cost constraints.

References

- Publications and books
  - FAO, 2015. World Programme for the Census of Agriculture 2020. FAO Statistical Development Series 15., Vol. 1. Rome
  - Global Strategy to improve agricultural and rural statistics., 2016. Hanbook on Agricultural Cost of production Statistics: Guide for data collection, compilation and dissemination., Rome, Italy
  - Global Strategy to improve agricultural and rural statistics., 2015. Hanbook of Master Sampling Frame for Agricultural Statistics: Frame development, Sample design and Estimation., Rome, Italy
  - Global Strategy to improve agricultural and rural statistics., 2012. Action Plan, Rome, Italy
  - Pascal Ardilly, 2006, les techniques de sondage. PARIS: Editions TECHNIP, 675 p

- Working document and web-references
  - Global strategy, 2015. Linking Area and List Frames in Agricultural Surveys. Technical Report Series GO-08-2015.
  - Cristiano, F., Leite, A., Ospina, R., et al. 2016. workshop on master sampling frame for agriculture surveys, Harare, Zimbabwe. UNECA
  - World Bank., 2010, Integrating Agriculture into National Statistical System, Workshop

Thank You