The Indirect Sampling as a General Approach for Defining Unbiased Sampling Strategies for integrated Agricultural Surveys

Pietro Gennari[1], Piero Demetrio Falorsi[2] and Clara Aida Khalil[2]
[1]Director Statistics Division, FAO, Roma, Italy
[2] FAO, Roma, Italy
Corresponding author: Pietro Gennari, e-mail: Pietro.Gennari@fao.org

Abstracts

The Global Strategy to Improve Agricultural and Rural Statistics, endorsed at the United Nations Statistical Commission in February 2010, underlines the need to ensure the consistency and the integration of agricultural statistics into national statistical systems, allowing building agricultural statistics in which the information on land parcels, households and farms are interlinked. The sample strategy, presented in this paper, achieves this strategic objective and simultaneously provides consistent statistics on the environmental, social and economic dimensions of agriculture. The methodological approach extends the use of indirect sampling to the case of producing integrated estimates on three target populations. The proposed techniques are quite flexible and may be tailored to the different informative contexts which characterize the production of agricultural statistics in developing countries. Furthermore, under quite general conditions, they allow to produce unbiased statistics, overcoming the majority of the problems caused by imperfect sampling frames.

Keywords: Consistency of estimates, Generalized Weight Share Method , Calibrated Estimates, integrated estimates.

1. Introduction

The Global Strategy to Improve Agricultural and Rural Statistics (GS), endorsed at the United Nations Statistical Commission in February 2010, is a comprehensive programme of statistical capacity development which aims at strengthening the availability and quality of agricultural statistics in developing countries (FAO, 2011).
This paper aims at describing the first results of a research project, conducted within the research program of the GS, and illustrates the main elements and properties of an unified survey sampling strategy which ensures at the same time the consistency of survey estimates and the correction of frame imperfections.
In order to ensure the consistency and the integration of agricultural statistics into the national statistical system, the GS stresses the need to build a Master Sampling Frame (MSF) where information on land parcels, households and farms are interlinked, thus allowing to simultaneously provide consistent and integrated statistics on the environmental, social and economic dimensions of agriculture (FAO, 2012). However, in many situations the construction of the MSF is not possible and, even when it is possible, it becomes rapidly outdated. In this paper we will show how the indirect sampling allows obtaining the integration overcoming the necessity of a MSF. Besides that, the indirect sampling techniques allow updating the MSF using the longitudinal links among units (Lavallé, 2007, ch 6). Indeed, resorting to this approach, the links among the units of the different target populations are built only during the data collection of the survey. Thus, there is no need of knowing the links for all the population units, as foreseen in the approach which base the integration on the availability of a MSF. Furthermore, the computation of integrated estimates, obtained through a common weighting technique on rectangular data set, is straightforward.
The paper is organized as follows. The informative context is introduced in section 2. The sample selection and a direct estimator are illustrated in sections 4 and 5,

respectively. The consistency is studied in section 6. The main results are summarized in section 7.

2. Informative context

Let ${}_jU$ be the unknown *j*-th target population of interest at the *current* time *t*, being *j*=1 for rural households, *j*=2 for farms and *j*=3 for land parcels. For a generic set *B* let $N_B = \#(B)$ denote the number of its element; thus, $N_{{}_jU}$ indicates the size of the population ${}_jU$.

Let ${}_jk$ denote the generic unit of population ${}_jU$, being ${}_jk = 1,...,N_{{}_jU}$, where ${}_1k$, ${}_2k$ and ${}_3k$ indicate respectively a household, a farm and a land parcel. The unit ${}_jk$ may viewed as a cluster, $U_{{}_jk}$, of $N_{U_{{}_jk}}$ elemental units ${}_j ki$ ($i = 1,...,N_{U_{{}_jk}}$).

Here below we assume that each variable of interest is **related** to only one of the target populations: e.g.: *employment status* is linked to the units of ${}_1U$; while, *maize production* is related to those of ${}_2U$.

Each complex unit has a single **reporting unit** which can provide the information for the whole unit. In the context here considered the reporting units are respectively: the head of the household, the farm holder and the farm holder in which the land parcel is located.

The variable of interest ${}_j y$, related to the population ${}_jU$, should be collected from the reporting units of the same population; but, in some situations, it is possible to collect the information from the reporting units of an alternative population. For instance, the *employment status*, related to ${}_1U$, should be measured through the households; however the information can be collected by asking to the farm holders (the reporting units of a different population) how many people have worked in the farm during the reference week; obviously, the latter measurement could be somehow biased.

Let ${}_j y_{{}_jk}$ be the value of the variable of interest ${}_j y$, (related to the population ${}_jU$) measured on the unit ${}_jk$ of the same population and let

$$ {}_jY = \sum_{{}_jk \in {}_jU} {}_j y_{{}_jk} = \sum_{{}_jk \in {}_jU} \sum_{i \in U_{{}_jk}} {}_j y_{{}_jki} \qquad (1) $$

be the *parameter of* interest.

Let ${}_jA$ be the sampling frame of the population ${}_jU$. The frames are usually built at the previous Census referred to time t_0, with $t_0 < t$. In the inter-censual period the frames become progressively outdated. Due to the temporal distance, the unit in the frame may differ from the corresponding unit at the current time *t* (e.g.: a farm can be divided or two people can form a new household).

A vector of auxiliary variables is available for each unit in the frame. This vector contains information of different nature: the census values of some target variables or the sample design variables (e.g. the stratification variables, the Primary Sampling Unit code and the Census Enumeration Area code).

3. Sampling

Alternative sampling schema may be proposed for producing consistent and unbiased inference on the parameters of interest. Each schema may be viewed as an ordered chain of samples.

The chain starts by selecting a direct sample of one of the target populations and then by observing the other populations by an indirect sampling. Conceptually it is possible to start the selection either from the households or from the farms or from the land parcels. Each choice has pros and cons that must be thoroughly examined in the

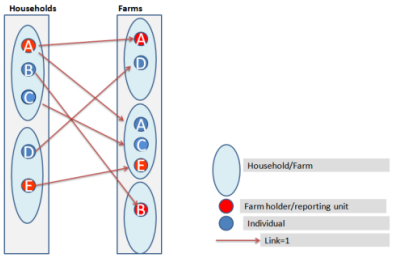
specific country informative context.

Starting from households

Let us consider below the case in which the selection starts from the households frame ${}_1A$. One possible design is the following: (1) Selecting a sample of Census Enumeration Areas (EAs) adopting a well known multistage stratified sampling design. (2) Making a census of all existing households in the sampled EAs.

Here below this schema is symbolized as ${}_1A \overset{EA}{\underset{dir}{\rightarrow}} {}_1S$. All the people of a sampled household are observed.

The subsequent sample, ${}_2S$, in the chain, finalized at individuating the farms, is done by indirect sampling. All the farms having people of ${}_1S$ as workers (either as employees or operators) are surveyed. In this way an indirect sampling ${}_2S$ (of unknown dimension ${}_2n = \#({}_2S)$) of the current population, ${}_2U$, of the farms is observed. In this indirect sampling process we observe the linking variables $l_{{}_1ki,\,{}_2ki'}$, in which: $l_{{}_1ki,\,{}_2ki'} = 1$ if the individual ${}_1ki$ is the individual *i'* working in the farm ${}_2k$, and $l_{{}_1ki,\,{}_2ki'} = 0$ otherwise. An example of this process is given in picture below, in which the ovals represent households or farms and the circles individuals; the first sample ${}_1S$ of two households is linked with three farms, where the individual A is the holder of the first farm and works in the second one.



In the last sample of the chain, all land parcels of the farms in ${}_2S$ are surveyed. Thus, an indirect sample ${}_3S$ (of unknown dimension ${}_3n = \#({}_3S)$) of the current population, ${}_3U$, of the land parcels is observed. In this case, the land may be considered as a variable related to the farm. The linking variables are defined as before. Therefore, the whole sequence of sample chain may be represented as

$${}_1A \overset{EA}{\underset{dir}{\rightarrow}} {}_1S \overset{person}{\underset{indir}{\rightarrow}} {}_2S \rightarrow {}_3S\,. \qquad (3.1)$$

Starting from land

A possible sample chain is the following. (1) Selecting a sample of EAs. (2) Making a census of all existing land parcels in the sample EAs, thus obtaining the sample ${}_3S$. (3) Creating an indirect sample of a farms, ${}_2S$, by considering only the farms which have their headquarter located in the area of the selected sample of land parcels. (3) Forming a sample of households by considering all the households of the farm workers in ${}_2S$. This chain may be represented as

$${}_3A \overset{EA}{\underset{dir}{\rightarrow}} {}_3S \overset{farm\,headquarter}{\underset{indir}{\rightarrow}} {}_2S \overset{worker}{\underset{indir}{\rightarrow}} {}_1S\,.$$

Note that the indirect sampling schema ${}_3S \overset{farm\,headquarter}{\underset{indir}{\rightarrow}} {}_2S$ is that proposed in the Area Frame sampling for the *Open Segment Estimator* (Faulkenberry and Garoui,

1991). An alternative chain may be defined if all the farms which have land in ${}_3S$ are surveyed. If the link ${}_{l_{3k,\,2k}}$ are identified as the proportion of the land parcel ${}_3k$ with respect to the farm ${}_2k$, the classical design proposed for the *Closed Segment Estimator* is obtained. The latter chain may be symbolized as

$$ {}_3A \overset{EA}{\underset{dir}{\rightarrow}} {}_3S \overset{prop.land}{\underset{indir}{\rightarrow}} {}_2S \overset{worker}{\underset{indir}{\rightarrow}} {}_1S . \tag{3.2} $$

Starting from farms with multiple frames
The process may start from the farms. In this context, a multiple frame (or a dual frame) approach is often adopted. A possible chain is the following. (1) Selecting a sample of farms ${}_2S$ by a multiple frame design. (2) All the land of the farms in ${}_2S$ is surveyed. (3) Forming a sample of households by considering all the households of the farm workers in ${}_2S$. This chain may be represented as

$$ {}_{multiple\,2}A \overset{mult}{\underset{dir}{\rightarrow}} {}_2S \overset{worker}{\underset{indir}{\rightarrow}} {}_1S \quad \text{and} \quad {}_2S \rightarrow {}_3S . $$

where ${}_{multiple\,2}A$ is the frame that may be identified by the union of different frames existing on the population ${}_2U$.

4. Estimation
The unbiased estimate of the totals ${}_jY$, may be obtained by the Direct Generalized Weight Share Method (DGWSM) estimator, as

$$ {}_j\hat{Y} = \sum_{{}_jk \in {}_jS} w_{{}_jk}\, {}_jy_{{}_jk} = \sum_{{}_jk \in {}_jS} w_{{}_jk} \sum_{i \in U_{{}_jk}} {}_jy_{{}_jki} \tag{4.1} $$

in which the weights ${}_{w_{jk}}$ are defined taking into account the specific sample chain. If the sample chain is the (3.1), then the sample weights are expressed by

$$ w_{{}_jk} = \begin{cases} 1/\pi_{{}_1k} & \text{if } j = 1 \\ \sum_{{}_1k \in {}_1S} \sum_{i \in {}_1k\,U} \dfrac{L_{{}_1ki,\,{}_jk}}{\pi_{{}_1k}\, L_{{}_jk}} & \text{if } j = 2 \end{cases} , \tag{4.2} $$

being $L_{{}_1ki,\,{}_jk} = \sum_{i' \in {}_jk\,U} l_{{}_1ki,\,{}_jki'}$, $L_{{}_jk} = \sum_{{}_1k \in {}_1U} \sum_{i \in {}_1k} l_{{}_1ki,\,{}_jki}$ where this information is directly collected from the sampled units.
The estimates ${}_3\hat{Y}$ may be directly obtained with the weights ${}_{w_{2k}}$ by summing the lands characteristics collected over the farms.
The weights expressed by (4.2) allow producing unbiased estimated for the population ${}_1U$ if the following condition holds:

Condition 3.1. The sampling frame ${}_1A$ allows for a full coverage of the population ${}_1U$; in symbols:

$$ {}_1U = \bigcup_{b=1}^{B} {}_1U_a , \tag{4.4} $$

in which ${}_1U_a$ denotes the subpopulation of households that may be identified making a census of the *b*-th EA recorded in the frame ${}_1A$.

Furthermore, using the results given in Lavallée (2007, ch. 4) it is possible to show that for the sample chain (3.1), the (4.1) - with the weights expressed by (4.2) - is unbiased for the population ${}_2U$ if the following condition holds.

Condition 3.2. *Each farm worker/holder* in U_2 *must have at least one link with a person in* ${}_1U$ *, in symbols:*

$$\sum_{{}_1k\in{}_1U}\sum_{i\in U_{1k}} l_{1ki,\,{}_2k} > 1 \quad \forall {}_2k \in U_2 \,. \tag{4.5}$$

An alternative expression of estimator (4.1) is

$${}_j\hat{Y} = \sum_{{}_1k\in{}_1S}\sum_{i\in{}_1k} d_{1k}\; z_{1ki} \tag{4.6}$$

in which

$$d_{1k} = 1/\pi_{1k}\,, \quad z_{1ki} = \begin{cases} {}_1y_{1k} & \text{if } j=1 \\ \sum_{{}_jk\in{}_jU}\sum_{i'\in{}_jkU} l_{1ki,\,{}_jki'}({}_jy_{{}_jk}/L_{{}_jk}) & \text{if } j=2 \end{cases}. \tag{4.7}$$

This expression of the estimator is interesting from an operational point of view, since it shows the possibility to build an unique data set related to the households, in which: (*i*) the sampling weights (d_{1ki}) are those defined in the sampling on the households; (*ii*) the variables values are those original, ${}_1y_{1k}$, if the variable is related to the population of households; (*iii*) the variables values are the transformed ones, ${}_1z_{1k}$, if the variable of interest is related to the populations of farms and land parcels. These variables take into account the links existing with the households.
As far as concerns, the other sample chains, we note that the extension of the estimator to the sample chain (3.2) is straightforward, while the extension to the chain (3.3) has to be based on the Generalized Multiplicity-Adjusted Horvitz Thompson Estimator (Singh and Mecatti, 2011).

5. Consistency

By (4.6), we note that the estimator (4.1) achieves the *internal consistency* of the estimates with respect to the three target populations. Furthermore if a vector, ${}_j\mathbf{x}_{jk}$, of auxiliary variables is available for the unit ${}_jk$ of the selected sample ${}_js$, the *Generalized Weight Share Estimator* (GWSE) also assures the coherence of the estimates with known benchmark totals ${}_j\mathbf{X} = \sum_{{}_jk\in{}_jU}\,{}_j\mathbf{x}_{jk}$ derived from some external sources.

Considering the sample chain (3.1), the GWSE is given by

$${}_j\hat{Y}_{greg} = \sum_{{}_1ki\in{}_1S}\; cal\, d_{1k}\; z_{1k}\,, \tag{5.1}$$

in which

$$ {}_{cal}d_{1k} = d_{1k}[1+({}_{j}\mathbf{X} - {}_{j}\hat{\mathbf{X}})'(\sum_{1k\in {}_{1}S} {}_{j}\tilde{\mathbf{x}}_{1k}\ {}_{j}\tilde{\mathbf{x}}_{1k}/\pi_{1k})^{-1}]\,{}_{j}\tilde{\mathbf{x}}_{1k}\ , \qquad (5.2) $$

where

$$ {}_{j}\hat{\mathbf{X}} = \sum_{1k\in {}_{1}S} {}_{j}\tilde{\mathbf{x}}_{1k}\ d_{1k}\ ,\ {}_{j}\tilde{\mathbf{x}}_{jk} = \begin{cases} {}_{1}\mathbf{x}_{1k} & \text{for } j=1 \\ \sum_{jk\in {}_{j}U}\sum_{i'\in {}_{jk}U} l_{1ki,jki'}({}_{j}\mathbf{x}_{jk}/L_{jk}) & \text{for } j\neq 1 \end{cases}. $$

The GWSE estimator is *calibrated* in the sense it assures that the estimated totals of the auxiliary variables ${}_{j}\hat{\mathbf{X}}$ are benchmarked to the total ${}_{j}\mathbf{X}$ known from the external sources.

6. Conclusions

In this paper we have shown as indirect sampling may represent a unified approach for assuring the consistency of integrated agricultural statistics, and for dealing with frame imperfection. The approach is general, flexible and may be tailored to the different country contexts. It enable us to achieve integration overcoming the necessity of a MSF. It represents a methodological solution which covers as particular cases different methods proposed in literature for dealing with imperfect frames and *rare populations* (as the snowball sampling or adaptive sampling (Chaudhuri, 2010). Furthermore, as shown in Lavallée and Rivest (2012), indirect sampling could represent a generalization of the usual Petersen estimator, used in the context of coverage errors and multiple frames.

References

Chaudhuri A. (2010), Estimation with inadequate frames, in Benedtetti R. Bee. M., Espa G., Piersimeoni F., *Agricultural Survey Methods*,Wiley, N.Y.

FAO (2011), *The Global Strategy to Improve Agricultural and Rural Statistics*.

FAO (2012), *Guidelines for Linking Population and Housing Censuses with Agricultural Censuses*.

Faulkenberry and Garoui A., (1991), Estimating a Population Total Using an Area Frame, *Journal of the American Statistical Association*, Vol.86, No. 414.

Lavallée P. and Rivest L.P. (2012) Capture-recapture sampling and indirect sampling. *Journal of Official Statistics*, 28, 1-27.

Lavallée P. (2007), *Indirect Sampling*. Springer.

Särndal, C.E., Swensson, B., Wretman, J., (1992). *Model Assisted Survey Sampling*, Springer-Verlag.

Singh, AC, Mecatti, F (2011), Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys, Journal of Official Statistics, Vol 27, No 4, 2011, pp 633–650.