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, based on indirect sampling and on the generalized weight share method for estimation, is 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, it allows to produce unbiased statistics, overcoming the majority of the problems caused by imperfect sampling frames.

Keywords: Consistency of estimates, Weight Share Method, Calibrated 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). 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 However, in dynamic environments, the MSF becomes quickly outdated, and the survey enumerators may encounter an actual situation quite different from that represented in the MSF.

Although the MSF allows to establish an effective link between the different statistical units, important sampling and estimation problems still remain to be addressed even in a simplified context where a single multipurpose survey is used to collect and disseminate statistics for the three target populations. These problems can be tackled by adopting indirect sampling strategies (Lavallé, 2007) as a unified approach both for sampling and direct estimation.

This paper aims at describing the first results of a research project, conducted within the research programme 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. The paper is organized as follows. The informative context is introduced in section 2. Section 3 discusses the concepts of consistency and coherence. The sample selection and a direct estimator are illustrated in sections 4 and 5, respectively. The main results are summarized in
2. NOTATION AND BACKGROUND

2.1. Parameters of interest

Let \( U_j \) 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; in this way \( N_j \) indicates the size of the population \( jU \).

Let \( k_j \) denote the generic unit of population \( jU \), being \( k_{1j} \) for a household, \( k_{2j} \) for a farm and \( k_{3j} \) for a land parcel. In case of either an household or a farm, the unit \( k_j \) may be viewed as a cluster, \( U_{kj} \), of \( N_{kj} \) elemental units \( k_{ij} \) (\( i=1,...,N_{kj} \)).

For the sake of simplicity, here below we assume that each variable of interest is related to only one of the target populations; e.g., \( \text{employment status} \) is related to the units of \( 1U \); \( \text{maize production} \) is related to those of \( 2U \).

In order to take into account the relationship between a variable of interest and the related population, it is useful to label generic variable of interest with the index of population to which it is related; thus, let \( jy_{jk} \) be the value of the variable of interest \( jy \), (related to the population \( jU \)) measured on the unit \( jk \) of the same population and let

\[
jY = \sum_{jk \in jU} jy_{jk} \tag{2.1}
\]

be the parameter of interest.

If the variable \( jy_{jk} \) may be expressed as sum of the values of elemental units of \( U_{jk} \):

\[
jy_{jk} = \sum_{i \in U_{jk}} jy_{ki} \]

then the total \( jY \) may be expressed alternatively by:

\[
jY = \sum_{jk \in jU} \sum_{i \in U_{jk}} jy_{ki} \tag{2.1.b}
\]

2.2. The observational context

The \textit{reporting unit} is the person who can provide the information on the variable value \( jy_{jk} \) for the whole complex 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.

In order to minimize the risk of measurement error, if the parameter of interest may be expressed in the form (2.1.b), it is safer not to gather directly the aggregated value \( jy_{jk} \) from the reporting unit of the complex unit, while it is better trying to collect the information \( jy_{ki} \) from the elemental units \( k_{ij} \).

As a general rule, the variable of interest, \( jy \), has to be collected from the reporting units of the population to which the variable is related or, in case of parameter expressed by 2.1.b, from their elemental units.

However, in some situations, it is either possible or necessary to collect the information from the units (either reporting or elemental) of an alternative population.
Example 2.1
Let us consider for instance the case of employment status, a variable related to the population \( U \). As a general rule, this variable should be measured by a survey on households. Once individuated the household, the information of interest could be taken either by asking to each member if he/she has worked in the reference week (the safer solution) or by asking the aggregated value (how many people of the households have worked in the reference week) to the reporting unit.

However, if it is necessary to collect the information by means of a survey on farms, it can be done by asking to each farm worker (the elemental units of a farm) how many people in her/his household have worked in the reference week. The latter measurement may be affected by bias, deriving from the fact that the people of a given household may be related to more than one farm. In the example in the figure 2.1 below, the person A may be counted twice. As will be discussed in sections 4 and 5, this potential bias may be avoided by asking to each worker the number of farms in which the members of her/his households are involved.

![Figure 2.1: Relationship between households and farms](image)

2.3. Sampling frames
Let \( A \) be the sampling frame of the population \( U \) and let \( k \) denote the generic unit of it. The sampling frames related to households and farms are usually built at the previous Census referred to time \( t_0 \), with \( t_0 < t \). In the inter-censal period these frames become progressively outdated. Due to the temporal distance the unit \( k \) may differ from the actual unit at time \( t \); e.g.: a farm can be divided or two people can form a new household.
The frames related to land parcels are less affected by the problem of updating. A vector of auxiliary variables, $x_{jk}$, 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 and the identifiers of Enumeration Areas (EA)).

Let $A_{1.2.3}$ be the Master Sampling Frame (MSF), which is a data base containing the information recorded in some of the frames $jA$ ($j=1,2,3$) and the links among the units of the different populations. Formally, the links are expressed as indicator variables $l_{jk}$, which is a vector in which $l_{jk} > 0$ if there exists a link between the units $j$ and $k$ (being $j \neq k$) and $l_{jk} = 0$ otherwise.

A lot of good practices illustrating how Population Census can be used for collecting agricultural variables are described in FAO (2012, Chapter, 4). Three main different approaches are proposed

- Inclusion of agricultural items in the population and housing census (e.g. Albania, Nepal, Mozambique, ...).
- Extended agricultural module in the population and housing census (e.g. Burquina Faso).
- Conducting the population and the housing census jointly with the agricultural census (Canada).

The definitions of farm holder and that of head of rural households are different (FAO, 1986, 2010). The farm holder may be individuated making a specific screening to the head of a rural household. The farm holders represent a subset of the heads of rural households. The screening is performed in Burkina Faso, while in Mozambique the additional screening is avoided and thus for all statistical purposes the head of a rural household coincides with the farm holder.

The MSF may be built also on a sampling basis. In many countries the Living Standard measurement Study (LSMS, 2013) and the Labor Force Survey would represent a good basis for the MSF construction. The specific strategies that can guide the MSF construction may be based on the three approaches just described. Let $(1.2.3)_{j}A$ indicate the MSF built on a sample basis.

In order to assure the possibility to build sound inferences on the target populations $jU$, each unit in $(1.2.3)_{j}A$ should have a weight, $w_{jk}$, which should incorporate all the relevant information on the sampling process.

Theoretically, the sampling weight should guarantee that

$$E_p \left( \sum_{j=1,2,3} w_{jk} \right) = jU N,$$

(2.2)
denoting with $E_p(\cdot)$ the sampling expectation.

A less strong condition that, under proper sampling schema, could guarantee the possibility of building sound inference is

$$E_p\left(\sum_{j,k=1,2,3} w_{jk}^A\right) = jA N \quad (2.2b)$$

The use of these weights will be discussed in a subsequent release of this work.

### 3. CONSISTENCY, INTEGRATION AND COHERENCE

#### 3.1. First sight on the concepts of coherence

In statistics the concept of consistency refers to different ideas. As far the inference is concerned, a consistent estimator is one for which, when the estimate is considered as a random variable indexed by the number $n$ of items in the data set, as $n$ increases the estimates converge to the value that the estimator is designed to estimate. In the measurement setting (Allen and Yen, 2002; Cronbach et al. 2004), the internal consistency is used to assess the correlation of different statistics in measuring the same unknown concept. When considering the field of random variables (Tarpey and Flury, 1996), the concept of self-consistency is related to that of regression and describes how the knowledge of the value of a given variable improves the prediction with respect to another variable. None of the above definitions meets all the needs that have guided the creation of a specific line of research within the GS. A concept that meets the above needs, though in non formal way, may be found in natural language where the consistency is defined as the agreement or harmony between parts of something complex. Similarly, it may be useful to consider the concept of coherence as defined in the Quality Framework of the Australian Bureau of Statistics (ABS, 2013): “The fifth dimension of quality in the ABS is Coherence. Coherence refers to the internal consistency of a statistical collection, product or release, as well as its comparability with other sources of information, within a broad analytical framework and over time. The use of standard concepts, classifications and target populations promotes coherence, as does the use of common methodology across surveys”. In this definition consistency and coherency are synonymous.

As seen above, the concept of consistency may be somehow misleading since it is used with different meaning in many statistical contexts. For avoiding any doubt, here below we adopt the concept of coherence that in our setting seems to be less controversial.

#### 3.2. Examples of coherence for integrated agricultural statistics

The need of jointly examining phenomena related to different target populations, have driven the creation of a specific research line within the GS.

Given the complexity of the topics, in order to come closer to a definition of the concepts of interest is useful to start from a few examples. See for instance the following table that defines some of indicators to be produced for the GS. In the example the yellow rows are indicators that are related to households (rows: 1,2,3), the green rows (4,…,9) denote indicators related to farms and the orange rows (11,…,14) refer to land parcels. The Consumer Price Index (line 10 in red) is a special case, since conceptually involves a survey on households for defining the base weights of the aggregated consumption and a survey on the business (points of

---

purchase of goods and service,…) as far concerns the dynamic of prices.

### Example 3.1: Indicators for Agricultural Statistics

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>Demographics of agricultural and rural population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>Rural poor as a percent of total poor population</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Rural hungry as a percent of total poor population</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Productivity of crop production as measured by crop yields</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Livestock value added</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Change in components of livestock and poultry balances by species</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Change in productivity of capture fish production</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Change in components of fish balances</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Change in components of forestry balances</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Commodity price indexes</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>Productivity of crop production as measured by crop yields</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>Livestock value added</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>Change in components of livestock and poultry balances by species</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>Change in productivity of capture fish production</td>
</tr>
</tbody>
</table>

Another hypothetical example is the following, in which the mean income of the households and the mean investment of farms are cross-tabulated by the type of farm.

### Example 3.2: Cross tabulated statistics

<table>
<thead>
<tr>
<th>Type of farm</th>
<th>Household Mean Income</th>
<th>Mean Investment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>$\bar{Y}_{\text{Income}}$</td>
<td>$\bar{Y}_{\text{Investment}}$</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Both tables represent complex statistical products.

We note the following aspects with regard to the example 3.1:

- The different indicators can be derived by three independent surveys: one on the households, and the others on the farms and land parcels. This solution is reasonable if the indicators for the units of the different populations are unrelated.
- However, if the indicators are somehow related, the independency of the surveys could expose to the risk of non-coherence among indicators, that may emerge from a joint analysis of the different indicators although every indicator separately considered is formally correct.
- A good prerequisite for obtaining the coherence seems that of jointly observing the different population units, involved in the indicators. As for instance, if a farm is included in the survey, then all the households with workers in that farm have to be surveyed. However, the joint observation has to take into account the complex relationships between the statistical units of the different populations (see example in figure 2.1).

The example 3.2 evidences the need of having a unique survey in which the units of the different populations are jointly observed, indeed:

- The information on the type of farm has to be collected by the farm holder.
- However in case of a farm in which there are people from more one household, the farm holder cannot be able to furnish the data on the households income.
Conversely, the head of household can be not able to give accurate information on the types of farm in which the members of her/his household are occupied.

The joint observation needs to take into account the fact that a household may be involved in more than one farm and vice versa.

**A Numerical Example**

In order to clarify the context outlined in Example 3.2, i.e. the necessity of leading a unique survey in which the units from different target populations are jointly observed, a numerical example may be useful.

Let’s suppose to conduct a survey on the average income of rural households and on the average product of farms.

For this purpose, two samples are collected, the first containing two rural households and the second containing the three farms in which the members of the considered households are workers. If these two surveys are performed separately one from each other, it is not possible to determine, for instance, the average income of the households starting from the information related to the average product of the farms.

Conversely, assuming to know the links among the units of the target populations here considered (illustrated in Figure 2.3) we are able to cross-tabulate these information about the income and the production, allowing the conduction of a more coherent survey.

Table 3.3 shows, for instance, that the first household has an average income of 100, of which a 25% derives from the first farms, a 50% from the second and a 25% from the third farm.

<table>
<thead>
<tr>
<th>Table 3.3: Average Income and Average Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farm_1</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Household_1</td>
</tr>
<tr>
<td>Household_2</td>
</tr>
<tr>
<td>Average Production</td>
</tr>
</tbody>
</table>
3.3. Some definition of coherence for integrated agricultural statistics

Here below we try to give some first formal definitions of coherence for integrated agricultural statistics.

Definition 1. Coherence of estimates with respect to a function.
Let \( Y_j \) and \( Y_j' \) denote two population parameters and let \( \hat{Y}_j \), \( \hat{Y}_j' \) indicate the corresponding estimates. Let \( g(a,b) \) a real function on two real values \( a \) and \( b \). If the function computed on the population values is bounded within the close interval \([L,U]\) \( g(Y_j, Y_j') \), the coherence of the estimates with respect to \( g \) implies \([L,U]\) \( g(Y_j, Y_j') \Rightarrow [L,U]\) \( g(\hat{Y}_j, \hat{Y}_j') \).

Example 3.3. Alternative estimates of the same population parameter
Let \( Y_j = Y_j' \) and let \( g(Y_j, Y_j') = Y_j / Y_j' = 1 \Rightarrow [1,1] \subseteq g(Y_j, Y_j') \). Let \( \hat{Y}_j \) and \( \hat{Y}_j' \) be two alternative estimates of \( Y_j \). The coherence of estimates with respect to the function \( g \) implies \( g(\hat{Y}_j, \hat{Y}_j') = \hat{Y}_j / \hat{Y}_j' = 1 \).

Definition 2. Coherence of estimates with respect to benchmark auxiliary variables.
Let \( X_j \) a known vector of total of auxiliary variables for the population \( jU \) and let \( \hat{X}_j \) the corresponding vector of sampling estimates. This kind of coherence implies that
\[
\hat{X}_j = X_j. 
\]

Definition 3. Coherence with respect to linking variables.
The linking variables play a fundamental role. If there is a link between the unit \( j,k^* \) and the unit \( j,k \) of a different population, then the reverse link holds. In symbols:
\[
l_{j,k^*,k} > 0 \Rightarrow l_{j,k,k^*} > 0. 
\]

3.3. Some strategies for coherence
A prerequisite for having coherent statistics is to have data sets containing information enabling to set up the linkages among the units of the different populations. This condition may be achieved adopting different survey strategies. A brief description of some of the feasible strategies is reported in the work-plan of the research activity on Integrated Survey Framework for Agricultural and Rural Statistics.
In this document we focus on a strategy based on indirect sampling. The idea is to start a direct sampling the units of a given population (e.g. the households) and then find the links to the units of other target populations (e.g. farms and land parcels) by mean of indirect sampling.
For the units observed with direct sampling the estimates are then computed with the usual estimators, while the Generalized Weight Share Method (GWSM) is used for the populations observed through indirect sampling.
4. SAMPLING

Alternative sampling schema may be proposed for obtaining a simultaneous coverage of the three target population. Each schema may be viewed as an ordered sequence of samples. This sequence starts by selecting a sample \( S_j \) either from \( A_j \) \((j=1, 2 \text{ or } 3)\) or \( 1,2,3.A \) or \((1,2,3)S.A \).

In order to exemplify the description, let us consider below the case in which the selection starts from the households frame \( 1.A \).

4.1. Starting from the households

The sample \( 1.S \) may be individuated either by direct or indirect sampling. One possible design for the direct sampling may be the following:

1) Selecting a sample of Enumeration Areas (EA).
2) Making a census of all existing households in the sample EAs.

This schema is symbolized here below as

\[
1.A \xrightarrow{\text{EA}} 1.S. \tag{4.1}
\]

One possible design for the indirect sampling may be the following:

1) Selecting a direct sample of households \( 1.S \) recorded in \( 1.A \).
2) Making an indirect sampling process in which each person of \( 1.S \) is linked with her/his new household of the sample \( 1.S \).

This sampling procedure can be denoted as

\[
1.A \xrightarrow{\text{EA}} 1.S \xrightarrow{\text{person}} 1.S \tag{4.2}
\]

Two kinds of variables play a fundamental role in the indirect sampling:

- The linking variables \( l_{a,b} \) between the units of populations \( a \) and \( b \);
- And the sample indicators variables \( t_b \) denoting if the units of population \( b \) have been observed in the sample of population \( a \).

The process of longitudinal identification of the new households with this kinds of variables may be summarized as follows. All the individuals of \( 1.S \), are surveyed and all the individuals of their (new) household are surveyed too irrespective of the fact if they were included or not in the original sample of individuals. We can thus define the sample indicator variables \( t_{1,ki} \) and the linking variables \( l_{1,ki,j} \), being: \( t_{1,ki} = 1 \) if \( 1.\{k_i\} \in 1.S \) and \( t_{1,ki} = 0 \) otherwise; \( l_{1,ki,j} = 1 \) if the individual \( 1.\{k_i\} \) belongs to the new household \( 1.j \) (in the sample \( 1.S \)), and \( l_{1,ki,j} = 0 \) otherwise. An example of this process is given in the figure 4.1, in which the ovals represents the households and the circles the individuals; the original sample \( 1.S \) is constituted by two households who generate three new households. All the six individuals of the three new households are surveyed. In the figure, the individual with non-zero links are A,B,D,E.

The relationship \( 1.S \xrightarrow{\text{person}} 1.S \), identify longitudinal links since the units in \( 1.S \) are
referred to the time $t_0$ while those in $t_1$ are referred to the current time $t$.

**Figure 4.1 Indirect sampling of new households**

### 4.1.1. Indirect sampling of farms and land parcels

The subsequent sample, $t_2$, in the chain, finalized at individuating the farms, is done by indirect sampling and finally all the land parcels of the selected farms are surveyed. The indirect sampling mechanism is different according to the way in which the first sampling on households has been selected.

**Case 1: $t_1$ selected by direct sampling**

All the farms having people of $t_1$ as workers (either as employees or holders) are surveyed. In this way an indirect sample $t_2$ of the current population, $t_2U$, of the farms is observed. We can thus define the linking variables $l_{i,k_l,2k}$ in which $l_{i,k_l,2k} = 1$ if the individual $i$ works in the farm $k_l$ in $t_2$, and $l_{i,k_l,2k} = 0$ otherwise. The sample indicators variables $t_{2k_l}$ assume value 1 if the farm worker (or the farm holder) is in the first sample, $t_1$, and assume value 0 otherwise.

An example of this process is given in figure 2, in which the ovals represents the households or farms and the circles the individuals; the first sample $t_1$ of two households is linked with three farms, where the individual A works in the two farms.

**Figure 4.2: Indirect sampling of farms**
In the last sample of the chain, all land parcels of the farms in $S_2$ are surveyed. In this way an indirect sample $S_3$ (of unknown dimension $n = \#(S_3)$) of the current population, $U_3$, of the land parcels is observed. In this case, the linking variables $l_{ik_1, k}$ assume value 1 if the individual $k_1$ works in the farm $k$ (of $S_3$) in which the sampled land parcel is located.

Therefore, the whole sequence of sample chain may be represented as

$$1 \stackrel{A_{ki}}{\rightarrow} 1 \stackrel{\text{indir}}{\rightarrow} 2 \text{dir} \stackrel{\text{dir}}{\rightarrow} S \rightarrow 3.$$  (4.3)

Using GWSM estimator (see section 5) this process produces unbiased estimates of the parameters of interest for the households, the farms and the land parcels iff the following constraints are satisfied.

**Constraint 4.1.** Each farm worker/holder in $U_2$ must have at least one link with a person in $U_1$, in symbols:

$$\sum_{1 \in U_1} \sum_{i \in U_{1k}} l_{1ki, 2k} > 1 \quad \forall 2k \in U_2.$$  (4.4)

**Case 2: $S_1$ selected by indirect sampling**

All the farms having people of $\tilde{S}_1$ as workers (either as employees or operators) are surveyed. In this way an indirect sampling $S_2$ (of unknown dimension $n = \#(S_2)$) of the current population, $U_2$, of the farms is observed. We can thus define the linking variables $l_{k_1, k}$ in which $l_{k_1, k} = 1$ if the individual $k_1$ works in the farm $k$ in $S_2$, and $l_{k_1, k} = 0$ otherwise. The sample indicators variables $t_{ki}$ assume value 1 if the farm worker (or the farm holder) is in the first sample $\tilde{S}_1$, and assume value 0 otherwise.

An example of this process is given in figure 4.3.
Therefore, the whole sequence of sample chain may be represented as

\[ A_{1a} \rightarrow^S_{\text{dir}} 1S \rightarrow_{\text{pair}} 1S \rightarrow 2S \rightarrow 3S \]  

Using GWSM estimator this sample chain produces unbiased estimates of the parameters of for households, farms and land parcels iff the following constraints are respected.

**Constraint 4.2.** Each new household and new farm must have at least one link with a person in \( A_1 \), in symbols:

\[
\sum_{i \in A_k} \sum_{j \in A_{1k}} l_{i_1j,ik} > 1 \quad \forall i \in U_1 \\
\sum_{i \in A_k} \sum_{j \in A_{1k}} l_{i_1j,2k} > 1 \quad \forall 2k \in U_2.
\]  

Note that the constraints (4.6a) and (4.6b) are much more binding of the constraints 4.3, and thus with this survey strategy could be more complex to obtain unbiased estimates for all the target populations.

### 4.2. Starting from the land

It is possible to start the sampling process from the land parcels. A possible sample chain is the following:

1) Selecting a direct sampling of land segments.  
2) Making a census of all existing tracts in the sample segments, thus obtaining the sample \( S_3 \).  
3) Creating an indirect sample of farms \( S_2 \), by considering only the farms which have their headquarter located in the area of the selected sample of land segments. The linking variables are \( l_{3k,2k} = 1 \) if the headquarter of the farm \( 2k \) is located in the area of the land segment \( 3k \) and equals 0 otherwise. The sample indicators are \( t_{2k} = 1 \) for all the farms belonging to \( S_2 \).  
4) Forming a sample of households by considering all the households of the farm workers in \( S_2 \). The linking variables are \( l_{2k,1k} = 1 \) if the individual \( 1k \) is a farm worker (or a farm holder) of \( 3k \) and equals 0 otherwise. The sample indicators are \( t_{2k} = 1 \) for all the individuals belonging to the farms in \( S_2 \).

This chain may be represented as

\[ A_3 \rightarrow^S_{\text{dir}} 3S \rightarrow^S_{\text{indir, farm headquarter}} 2S \rightarrow^S_{\text{indir, worker}} 1S. \]  

Note that the indirect sampling schema \( S_3 \rightarrow 2S \) is that proposed in the Area Frame sampling for the Open Segment Estimator (Faulkenberry and Garoui, 1991).
The step 3 is illustrated in figure 4.4. Only the farms 1, and 2 which have headquarters in the selected segments are included in the sample.

![Figure 4.4 Indirect sampling of farms - Open segment estimator](image)

The step 4 is illustrated in figure 4.5.

![Figure 4.5: Indirect sampling of households from farms](image)

Note that a sampling process that drives to well-known Weighted Segment Estimator for the area frame may be obtained if the step 3 above described is substituted by the following step 3a

3a) Creating an indirect sample of farms, $2S$, by considering all the farms which have land covering a selected tract. The linking variables $l_{3ki,2k}$ are equal to the proportion of the area of the quarter $3ki$ with respect to the arable area of the farm $2k$. The sample indicators are $t_{2ki} = 1$ for all the farms belonging to $2S$. This indirect sampling schema is described in figure 4.4a below.
The latter chain may be symbolized as

\[ 3A \xrightarrow{\text{EA}} 3S \xrightarrow{\text{prop.land.indir}} 2S \xrightarrow{\text{worker.indir}} 1S. \quad (4.8) \]

Using GWSM estimator the above sample chain produce unbiased estimates of the parameters of for households, farms if the following constraint is satisfied.

**Constraint 4.3.** Each new household and each new farm must have at least one link with a person in \(1A\), in symbols:

\[
\begin{align*}
\sum_{j \in A} \sum_{i \in A} l_{k,i,j} &> 1 \quad \forall i \in U_1 \\
\sum_{j \in A} \sum_{i \in A} l_{k,i,j} &> 1 \quad \forall j \in U_2.
\end{align*}
\]

\[
(4.9a) \quad (4.9b)
\]

### 4.3. Starting from the farms

The process may start from the farms.

Two possible sample chains are the following. In the second chain in which the longitudinal links between the old and the new farm is individuated by the farm workers or by the farm holders (see figure 4.6.):

\[
\begin{align*}
2A \xrightarrow{\text{EA}} 2S \xrightarrow{\text{worker.indir}} 1S \quad \text{and} \quad 2S \rightarrow 3S. \quad (4.10) \\
2A \xrightarrow{\text{EA}} 2S \xrightarrow{\text{worker.indir}} 2S \xrightarrow{\text{worker.indir}} 1S \quad \text{and} \quad 2S \rightarrow 3S. \quad (4.11)
\end{align*}
\]
Alternative sample chains may be identified if the longitudinal links between the old and the new farm are defined considering the changes in the land tracts.

The constraint conditions assuring the estimates are unbiased may be easily derived.

4.4. Some preliminary conclusion

Conceptually it is possible to start the selection either from the households or the farms or the land. Each choice has pros and cons that must be thoroughly examined in the specific country informative context. The approach here proposed is quite flexible and may be adopted for the particular country case.

Note that the methodological solution here described covers as particular cases different methods proposed in literature for dealing with

✓ imperfect frames
✓ 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.

5. ESTIMATION

5.1. Direct estimator

The unbiased estimate of the totals \( \hat{Y}_j \), may be obtained by the Direct Generalized Weight Share Method (DGWSM) estimator, as

\[
\hat{Y}_j = \sum_{k \in S} w_{jk} Y_k,
\]

in which the weights \( w_{jk} \) are defined taking into account the specific sample chain.

If the sample chain is

\[
1 \xrightarrow{A_{1i}} 1 S \xrightarrow{A_{2i}} 2 S \xrightarrow{A_{3i}} 3 S,
\]

then the sample weights are expressed by

\[
w_{jk} = \begin{cases} 
1/\pi_{jk} & \text{if } j = 1 \\
\frac{1}{\sum_{l \in U} \sum_{i \in l} t_{1ki} l_{1ki, jk} \frac{1}{\pi_{1ki} L_{jk}}} & \text{if } j \neq 1,
\end{cases}
\]

being \( \pi_{1ki} \) the inclusion probabilities in \( 1 S \) of the individual \( 1 ki \). Note that, being \( t_{1ki} = 0 \) if \( 1 ki \not\in 1 S \), it is not necessary to know the inclusion probabilities of the individuals not included in the first sample \( 1 S \).
Lavallée P. (2007, ch. 4) shows that this estimator is unbiased.

An alternative expression of estimator (5.1) is

\[ j \hat{Y} = \sum_{i \in 1} \sum_{k \in 1} d_{i,ki} \bar{z}_{i,ki} \]  

(5.3)

in which

\[ d_{i,ki} = \frac{1}{\pi_{i,k}}, \quad \bar{z}_{i,ki} = \begin{cases} \frac{1}{j} \sum_{j,k \in J} \sum_{i \in j,k} l_{i,ki,j,k} (jy_{j,k}/L_{j,k}) & \text{if } j = 1 \\ 0 & \text{if } j \neq 1 \end{cases} \]  

(5.4)

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:

- The sampling weights (\( d_{i,ki} \)) are those defined in the sampling on the households.
- The variables values are those original, \( 1y_{i,k} \), if the variable is related to the population of households.
- The variables values are the transformed ones, \( 1z_{i,k} \), 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.

With a data set organized in this ways, the production of the survey estimates, becomes an extremely straightforward operation.

Finally, we note, that this kind of estimator, allows respecting the condition of coherence (3.1). Consider for instance the case illustrated in the example 2.1. It is easy to demonstrate that the estimate of a variable of interest using the data collected on the units of the population to which the variable is related coincides with the alternative estimate which may be obtained if the variable of interest is collected from the units of alternative populations.

The extension of the estimates to the other sample chains is straightforward.

### 5.2. Calibrated estimator

This kind of estimators allows respecting the condition of coherence 3.2, and thus for, the survey estimates are benchmarked with the known totals of auxiliary variables every population of interest. We illustrate this kind of estimators, in the case of the Generalized REGression estimator (\( greg \)) (Särndal et al., 1992).

The form of the estimator depends on the sample chain used for collecting the data. In order to exemplify the description, let consider the sample chain

\[ 1A_{e1} \rightarrow 1S \rightarrow 2S \rightarrow 3S. \]

The extension to other sample chains is straightforward and will be developed in a further release of this paper.

In the case of the above chain, the \( Greg \) form of the Generalized Weight Share Estimator (GGWSE) is given by:
\[ j \hat{Y}_{\text{reg}} = \sum_{j \in j_S} \text{cal} d_{jk} \cdot z_{jk}, \]  

(5.5)

in which

\[ \text{cal} d_{jk} = d_{jk} [1 + (j \hat{X} - j \hat{X}) (\sum_{k \in 1S} j \hat{x}_{jk} j \hat{x}_{jk} / \pi_{jk})^{-1}] j \hat{x}_{jk}, \]  

(5.6)

where

\[ j \hat{X} = \sum_{j \in j_A} j x_{jk}, \quad j \hat{X} = \sum_{j \in j_S} j \hat{x}_{jk} d_{jk}, \]

\[ j \hat{x}_{jk} = \begin{cases} j x_{jk} & \text{for } j = 1 \\ \sum_{j \in j_A} \hat{L}_{jk} j (j x_{jk} / L_{jk}) & \text{for } j \neq 1 \end{cases}, \]  

(5.7)

We note the following aspects:

✓ The GGWSE estimator is calibrated in the sense it assures that for each of the target populations (households, farms and land parcels) the estimated totals of the auxiliary variables \( j \hat{X} \) are benchmarked to the total \( jX \) known from the frames.

✓ The GGWSE estimator assures the condition of coherence 3.1.

✓ For its calculation it is necessary to know the links from the sample and the known totals of the auxiliary derived independently from the different frames \( jA \).

✓ It is extremely easy to be used. Indeed,

- The calibrated sampling weights \( \text{cal} d_{jk} \) are those defined in the sampling on the households.
- The variables values are those original, if the variable is related to the population of households.
- The variables values are the transformed ones, 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.

Finally, we note that, in case independently selected samples, the consistency for a target variable may be obtained adopting a calibrated estimator, similar to (6), in which one of the benchmark totals is represented by a convex combination of the different direct estimates of the total of interest (Knotterus, 2007).

6. PRELIMINARY 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 very general and the different methods proposed in literature for dealing imperfect frames (as the snowball sampling or adaptive sampling (Chaudhuri, 2010)) may be viewed as particular case of indirect sampling. Also the
traditional estimators for Area Frame in agricultural statistics, as the Open Segment Estimator and the Closed Segment Estimator (Faulkenberry and Garoui, 1991), may be viewed as a particular expression of the DGWSM estimator; but this topic is not demonstrated in this paper. Furthermore, as shown in Lavallée and Rivest (2012), the indirect sampling could represent a generalization of the usual Petersen estimator, used in the context of coverage errors; indeed, with the joint use of indirect sampling and of the DGWSM estimator, it is possible to estimate also the totals and not only the size of the population. In this version of the report, the relevance of the MSF has not been studied in depth. Here we stress it highly improves the quality of the sampling, since the enumerators can be supported by the knowledge of the previous links recorded in the data base.

REFERENCES

FAO (1986), Food and Agricultural Statistics in the context in the Context of a National Information System.
FAO (2010), A System of Integrated Agricultural Censuses and Surveys.