Master frames for integrated and linked surveys

Verónica Boero* and Luis Ambrosio**
* Regional Office for Latin America and the Caribbean
Food and Agriculture Organization - United Nations.
Santiago. Chile
Veronica.Boero@fao.org
** Universidad Politécnica de Madrid. Dpt. Economía, Estadística y Gestión de Empresas
Madrid. España
luis.ambrosio@upm.es

DOI: 10.1481/icasVII.2016.f36e

ABSTRACT

To improve agricultural and rural official statistics, FAO designs and carries out action plans. The World Plans with specific focus have a long history: they began in 1930 and the last focuses on integrating agricultural censuses and agricultural surveys within the National Statistical System (NSS). The most recent Global Strategy [FAO (2011, 2012)] focuses on developing master sampling frames that are integrated with the NSS and allow the linkage of the farm as an economic unit to the household as a social unit and both to the land as an environmental unit [FAO (2015)]. Two keywords in this last plan are 'integration' and 'linkage'. 'Integration' refers to the use of the same sampling frame and related materials in multiple surveys, as well as the same concepts, survey personnel, and facilities. 'Linkage' is the basis for analysing the relationships among the economic, the social and the environmental dimensions of sustainable development.

FAO (1996, 1998, 2015) and the United Nations Statistical Division (UNSD,1986, 2008) have elaborated guidelines to assist countries in planning and implementing agricultural and household surveys, respectively. The central topic of these guidelines is the development and maintenance of master sampling frames. In this paper we focus on the integration of a dual sampling frame for agriculture with a sampling frame for households to build a multiple sampling frame that allows the required linkage among reporting units. We apply this strategy in three Latin America countries. We consider multiple frame regression estimators, highlighting its usefulness to integrate register and survey data and for combining data from independent surveys.

Keywords: Multiple overlapping frames, regression estimators, integrating register and survey data, combining data from independent surveys.

1. Introduction

There is consensus in the scientific community about the multidimensional (economical, social, and environmental) nature of sustainable development, and the Global Strategy to improve agricultural and rural official statistics [FAO (2011, 2012)] lays great stress on using a sampling frame that allows for the linkage of the farm as an economic unit, to the household as a social unit, and both to the land as an environmental unit. Multiple sampling frames meet this requirement and are cost-efficient.

Sampling strategies based on multiple overlapping frames have deserved a notable attention in last years, as a tool to deal with nonsampling errors: undercoverage, nonresponse, and measurement errors [Lohr (2011)]. In this paper we follow this sampling strategy for integrating agricultural and household surveys. Focus is on the linkage among farms, households and parcels. We structure the paper as follow. In section 2 we present the sampling strategy and its application in three Latin America countries. In section 3 we consider the multiple overlapping frame estimators proposed in the literature. In section 4 we consider multiple frame regression estimators, highlighting its usefulness to integrate survey and register data. Section 5 is about concluding remarks.

2. Integration of agricultural and household sampling frames

The sampling frames recommended by FAO and UNSD guidelines are dual frames, with an area component and a list component. The area frame ensures completeness, accuracy and up-to-datedness of the master frame: it is well established in the literature [Fecso et al. (1986), Faulkenberry and Garoui (1991), Vogel (1995), Ambrosio and Iglesias (2014)]. In agricultural surveys, the list contains the largest farms and contributes to improve the area sample accuracy. Census enumeration areas are used in household surveys as Primary Sampling Units (PSUs) and a list is elaborated within selected PSUs and is used to select the household sample.

We integrate the agricultural sampling frame and the household sampling frame in a unique multiple sampling frame. This multiple frame provides farms to observe economic variables: acreage and crop yields, livestock production, aquaculture and forestry. It provides also households to observe social variables: household composition, living conditions, employment, income, food and hunger, poverty, or inequality. And it provides parcels to observe environmental variables: soil degradation, water consumption for irrigation, or the quantity used of chemical fertilizers, herbicides, pesticides and fungicides.

Sampling a population with multiple overlapping frames

We use P to refer either the farms population, $F=\left\{ f \middle| f=1,2,\cdots,F \right\}$, the parcels population, $L=\left\{ l \middle| l=1,2,\cdots,L \right\}$, or the households population, $H=\left\{ h \middle| h=1,2,\cdots,H \right\}$. We assume that each population unit, $\mathrm{jj}\in \mathrm{P}$ $\left(\mathrm{jj}\equiv f,h,l \right)$, is associated to at least one sampling unit in the multiple frame

$\mathrm{I} = \left\{ i = 1, 2, \cdots, A^q; q = 1, 2, \cdots, Q \right\}$, where A^q denotes both, the generic single frame q and the number of sampling units, and Q is the number of single frames. We define the indicator variable $I_{ij}^q = 1$ if the population unit $jj \in \mathrm{P}$ is associated to the sampling unit $i \in A^q$, and $I_{ij}^q = 0$ otherwise $\left(jj \equiv f, h, l \right)$.

The sample

We select a set of samples $\left\{ \mathrm{S}^q; q = 1, 2, \cdots, Q \right\}$ independently from each frame A^q, using a sampling scheme that associates to sampling unit $i = 1, 2, \cdots, A^q$ an inclusion probability π_i^q. From the standard dual frame for agricultural surveys, where A^1 is an area frame with $i = 1, 2, \cdots, A^1$ segments and A^2 is a list frame with $i = 1, 2, \cdots, A^2$ names of farms, we select independently a sample S^1 of segments and a sample S^2 of names. From the standard frame for household surveys, A^3, we select a sample S^3 of names, independent of S^1 and S^2.

As a result, we have: (i) a sample of parcels, $\mathrm{S}_L^1 = \left\{ l \in L \middle| i \in \mathrm{S}^1 \wedge I_{il}^1 = 1 \right\}$, where $I_{il}^1 = 1$ when the area, a_{il}, of parcel l within the segment $i \in \mathrm{S}^1$ is $a_{il} > 0$, (ii) and a set of two partially overlapping samples of farms, $\left\{ \mathrm{S}_F^q; q = 1, 2 \right\}$, where $\mathrm{S}_F^q = \left\{ f \in F \middle| i \in \mathrm{S}^q \wedge I_{if}^q = 1 \right\}$, where $I_{if}^1 = 1$ when the area, a_{if}, of the farm f within the segment $i \in \mathrm{S}^1$ is $a_{fi} > 0$, and $I_{if}^2 = 1$ when the name $i \in \mathrm{S}^2$ is associated to the farm f, (iii) and a sample of households $S_H^3 = \left\{ h \in H \middle| i \in \mathrm{S}^3 \wedge I_{ih}^3 = 1 \right\}$, where $I_{ih}^3 = 1$ when the name $i \in \mathrm{S}^3$ is associated to the household h.

Country examples

We study the case of three Latin American countries: Guatemala, Costa Rica and Ecuador. In these countries, there is a dual sampling frame for agricultural surveys. The kind of limits used to define sampling units differs among countries: limits are geometrics in Guatemala and Ecuador, while identifiable physical boundaries are used in Costa Rica. The area frame, A^1, has 190100 segments in Guatemala [Ambrosio (2013), FAO (2015)], 352254 segments in Ecuador (Ambrosio, 2014) and 120326 segments in Costa Rica (Ambrosio, 2015).

The area frame is stratified into four strata, using the percentage of cultivated surface as stratification variable. The data source for stratification is a land use map in Guatemala and Ecuador and a geo-referenced agricultural census in Costa Rica. A target segment size is defined that varies among strata: in Guatemala it ranges from 6.25 hectares (cultivated surface bigger than 60% and small fields) to 100 hectares (cultivated surface lower than 20%), in Ecuador the range is from 9 to 576 hectares, and in Costa Rica the range is from 10 to 100 hectares. In Guatemala and Costa Rica S^1 has 1500 segments, and in Ecuador 5520 segments. The sample is allocated to strata according to Neyman's criterion, and five replicated samples are selected in each stratum.

The list frame, A^2, differs among countries according to available resources. In Costa Rica, there is a recent agricultural census and a list frame for each one of the main crops and animal' species is available (the bovine list frame has 31171 farms, and the porcine list frame has 14355 farms). In

Guatemala and Ecuador, the agricultural censuses are obsoletes and the number of list frames is reduced to the biggest farms in Ecuador and to the main animals' species in Guatemala.

An area sampling frame of enumeration areas (EA) with mapped, well-delineated boundaries is available for household surveys. In Guatemala the frame has 15511 EA with an average of 140 households by EA. The EA are stratified using available population figures, and a two-stage

sampling scheme is used to select the household sample, S^3. In the first stage, a sample of EA is selected with probabilities proportional to size (in Ecuador the sample size is 2586 EA for labour surveys, 1128 EA for surveys on living standard and 3411 EA for income surveys). In the second-stage, a list of household is updated within each EA in the first-stage sample and a sample of households (12 by EA) is selected with equal probabilities.

3. Multiple frame estimators

Typically, a population unit (e.g. a farm) is covered by two or more single frames (e.g., area and list frames) and, as a result, the weight estimator, $\hat{Y}_{\mathrm{P}} = \sum_{q=1}^{Q} \sum_{i=1}^{S^q} w_i^q y_i$, where $w_i^q = \frac{1}{\pi_i^q}$, is a biased estimator of the population total, Y_{P}. To see this, consider the population partitioned into $D = 2^Q - 1$ non-overlapping domains and $Y_{\mathrm{P}} = \sum_{d=1}^{D} Y_d$, where Y_d is the domain total, $d = 1, 2, \cdots, D$.

For dual frames, $Q = 2$, $\hat{Y}_{\mathrm{P}} = \sum_{q=1}^{2} \sum_{i=1}^{S^q} w_i^q y_i = \sum_{i=1}^{S^1} w_i^1 y_i + \sum_{i=1}^{S^2} w_i^2 y_i$, and $D = 2^2 - 1 = 3$. Domain $d = 1$ is the set of units covered only by A^1, domain $d = 2$ is covered only by A^2 and domain $d = 3$ is covered by both, A^1 and A^2. The population total is $Y_{\mathrm{P}} = \sum_{d=1}^{3} Y_d = Y_1 + Y_2 + Y_3$. Now, $\sum_{i=1}^{S^1} w_i^1 y_i$ is a unbiased estimator of A^1 total, which is domain $d = 1$ total plus $d = 3$ total, $Y_1 + Y_3$, and $\sum_{i=1}^{S^2} w_i^2 y_i$ is a unbiased estimator of A^2 total, which is $d = 2$ total plus domain $d = 3$ total, $Y_2 + Y_3$. Thus, $E\hat{Y}_{\mathrm{P}} = E\sum_{i=1}^{S^1} w_i^1 y_i + E\sum_{i=1}^{S^2} w_i^2 y_i = Y_1 + Y_2 + 2Y_3$ and the bias of $\hat{Y}_{\mathrm{P}}$ is $B\hat{Y}_{\mathrm{P}} = E\hat{Y}_{\mathrm{P}} - Y_{\mathrm{P}} = Y_3$.

A screening approach is followed in FAO (1996, 1998), where the single frames are pre-screened to remove overlap, so that domains with two o more frames are empty and, as a result, the weight estimator is unbiased: for dual frames, $d = 3$ is empty, $Y_3 = 0$, and hence $B\hat{Y} = 0$. However, screening operations are resource-consuming and a number of more cost-efficient alternatives can be found in the literature (Lohr, 2011). Cost-efficiency was the motivation of Hartley (1962, 1974) to propose first multiple frame estimators. Skinner and Rao (1996) and Lohr and Rao (2000, 2006) proposed pseudo-maximum likelihood multiple frame estimators. Bankier (1986) and Kalton and Anderson (1986) proposed standard single-frame estimators for multiple frame survey.

Adjusted-weight estimators

Most of these alternatives look for an adjustment, m_i^q, of the sampling weight w_i^q in such a way that using $\tilde{w}_i^q = m_i^q w_i^q$ instead of w_i^q, the adjusted-weight estimator $\hat{Y}_{\mathrm{P}} = \sum_{q=1}^{Q}\sum_{i=1}^{S^q} \tilde{w}_i^q y_i$ is unbiased. This can be achieved using for each frame and domain a fixed set of adjustment such as $\forall i \in d$, $m_i^q = m^{(q,d)}$, with the restrictions $m^{(q,d)} > 0$ (if domain d is not part of A^q, then $m^{(q,d)} = 0$) and $\sum_{q=1}^{Q} m^{(q,d)} = 1$ for $d = 1,2,\cdots,D$. The adjusted-weight estimator $\hat{Y}_{\mathrm{P}} = \sum_{d=1}^{D} \hat{Y}_d$, where $\hat{Y}_d = \sum_{q=1}^{Q}\sum_{i=1}^{S^q} \tilde{w}_i^q \delta_i(d) y_i = \sum_{q=1}^{Q} m^{(q,d)} \sum_{i=1}^{S^q} w_i^q \delta_i(d) y_i$ and $\delta_i(d) = 1$ if $i \in d$ and $\delta_i(d) = 0$ otherwise, is unbiased .

For dual frames, a fixed weight adjustment is: if $i \in (d=1)$ then $m_i^1 = m^{(1,1)} = 1$ and $m_i^2 = m^{(2,1)} = 0$, if $i \in (d=2)$ then $m_i^1 = m^{(1,2)} = 0$ and $m_i^2 = m^{(2,2)} = 1$ and if $i \in (d=3)$ then $m_i^1 + m_i^2 = m^{(1,3)} + m^{(2,3)} = 1$. The adjusted-weight estimator is $\hat{Y}_{\mathrm{P}} = \sum_{d=1}^{3} \hat{Y}_d$, where

$$\hat{Y}_1 = \sum_{q=1}^{2}\sum_{i=1}^{S^q} \tilde{w}_i^q \delta_i(1) y_i = \sum_{q=1}^{2} m^{(q,1)} \sum_{i=1}^{S^q} w_i^q \delta_i(1) y_i = m^{(1,1)} \sum_{i=1}^{S^1} w_i^1 \delta_i(1) y_i + m^{(2,1)} \sum_{i=1}^{S^2} w_i^2 \delta_i(1) y_i = \sum_{i=1}^{S^1} w_i^1 \delta_i(1) y_i$$

, $\hat{Y}_2 = \sum_{i=1}^{S^2} w_i^2 \delta_i(2) y_i$ and $\hat{Y}_3 = m^{(1,3)} \hat{Y}_3^1 + m^{(2,3)} \hat{Y}_3^2$, where $\hat{Y}_3^1 = \sum_{i=1}^{S^1} w_i^1 \delta_i(3) y_i$ and $\hat{Y}_3^2 = \sum_{i=1}^{S^2} w_i^2 \delta_i(3) y_i$.

Often, it is taken $m^{(1,3)} = m^{(2,3)} = \frac{1}{2}$ and, as a result, $\hat{Y}_3 = \frac{1}{2}\hat{Y}_3^1 + \frac{1}{2}\hat{Y}_3^2$.

Optimal estimators

Hartley (1962) proposes this other fixed set of adjustments: if $i \in (d=1)$ then $m_i^1 = m^{(1,1)} = 1$ and $m_i^2 = m^{(2,1)} = 0$, if $i \in (d=2)$ then $m_i^1 = m^{(1,2)} = 0$ and $m_i^2 = m^{(2,2)} = 1$ and if $i \in (d=3)$ then $m_{i,\theta}^1 = m_\theta^{(1,3)} = \theta$ and $m_{i,\theta}^2 = m_\theta^{(2,3)} = 1-\theta$, where $0 \le \theta \le 1$. The adjusted-weight estimator is $\hat{Y} = \sum_{d=1}^{3} \hat{Y}_d$, where $\hat{Y}_1 = \sum_{i=1}^{S^1} w_i^1 \delta_i(1) y_i$, $\hat{Y}_2 = \sum_{i=1}^{S^2} w_i^2 \delta_i(2) y_i$ and $\hat{Y}_3 = \theta \hat{Y}_3^1 + (1-\theta)\hat{Y}_3^2$, so that $\hat{Y}_{\mathrm{P}} = \hat{Y}_1 + \hat{Y}_2 + \theta \hat{Y}_3^1 + (1-\theta)\hat{Y}_3^2$. The value $\theta = \frac{1}{2}$ is often used and the estimator is internally consistent. However, the optimal value is $\theta_H = \frac{V\hat{Y}_3^2 + Cov(\hat{Y}_3^2, \hat{Y}_2) - Cov(\hat{Y}_3^1, \hat{Y}_1)}{V\hat{Y}_3^1 + V\hat{Y}_3^2}$ and changes with the survey variable, so that it is internally inconsistent. In practice, internal consistency requires that one set of weights be used to estimate all survey variables: Pseudo-maximum likelihood estimators are internally consistent (Lohr, 2011).

Single frame estimator

Kalton and Anderson (1986) propose an adjustment weight, which treats all observations as though they had been sampled from one frame: if $i \in (d=1)$, then $m_{i,S}^1 = 1$, if $i \in (d=2)$ then $m_{i,S}^2 = 1$and if $i \in (d=3)$then $m_{i,s}^1 = \frac{w_i^2}{w_i^1 + w_i^2}$ and $m_{i,s}^2 = \frac{w_i^1}{w_i^1 + w_i^2}$. If $i \in (d=3)$then $\tilde{w}_i^1 = \tilde{w}_i^2 = \frac{1}{\pi_i^1 + \pi_i^2}$. This estimator is internally consistent.

Multiplicity-adjusted estimators.

Singh and Mecatti (2011) and Mecatti and Singh (2014) propose to adjust for multiplicity the survey variable value, instead of the sampling weight. The multiplicity of a population unit, $j \in \mathrm{P}$ ($j = f,h,l$ and $\mathrm{P} = F,H,L$), is the number of sampling units, $m_j = \sum_{q=1}^{Q} m_j^q$, to which it is associated, jj ∈ P where $m_j^q = \sum_{i=1}^{A^q} I_{ij}^q$ is the multiplicity within A^q, where $I_{ij}^q = 1$ if the population unit is associated to the sampling unit $i \in A^q$, and $I_{ij}^q = 0$ otherwise. The population total is $Y_{\mathrm{P}} = \sum_{q=1}^{Q} \sum_{i=1}^{A^q} \tilde{y}_i^q$, where $\tilde{y}_i^q = \sum_{j=1}^{\mathrm{P}} \alpha_{ij}^q y_j$ is the multiplicity-adjusted value of the survey variable in the i^{th} sampling unit, where $\alpha_{ij}^q = \frac{I_{ij}^q}{m_j}$.

The weight multiplicity-adjusted estimator, $\hat{Y}_{\mathrm{P}} = \sum_{q=1}^{Q} \sum_{i=1}^{S^q} w_i^q \tilde{y}_i^q$, is unbiased and internally consistent.

Note that the adjustment, $\frac{1}{m_j}$, applies to the survey variable value, y_j, instead to the sampling weight, w_i^q, and it consists in sharing y_j among the number of sampling units to which $j \in \mathrm{P}$ is associated.

In terms of the population units, the multiplicity-adjusted estimator can be written as an adjusted-weight estimator, $\hat{Y}_{\mathrm{P}} = \sum_{q=1}^{Q} \sum_{j=1}^{S_{\mathrm{P}}^q} \tilde{w}_j^q y_j$ where $\mathrm{S}_{\mathrm{P}}^q = \left\{ j \in \mathrm{P} \middle| i \in \mathrm{S}^q \wedge I_{ij}^q = 1 \right\}$ is the set of population units associated to S^q and $\tilde{w}_j^q = \frac{1}{m_j} \sum_{i=1}^{S^q} w_i^q$. The size of $\mathrm{S}_{\mathrm{P}}^q$ is $\mathrm{n}_{\mathrm{P}}^q$.

Linkage

To ensure the required linkage between farms and households, we define the link $I_{fh} = 1$ if at least one person from the household $h \in H$ works for the farm $f \in F$ and $I_{fh} = 0$ otherwise. A parcel is linked to the farm which it belong and to the households through the linkage between farms and households: $I_{lf} = 1$ if l belongs to f and $I_{lf} = 0$ otherwise. If f is included in S_F^1 (see section 2), then the set of households $\mathrm{S}_H^1 = \left\{ h \in H \middle| f \in S_F^1 \wedge I_{fh} = 1 \right\}$ linked with f are included in the

household sample. If f is included in S_F^2, then the set of households $S_H^2 = \{h \in H \mid f \in S_F^2 \wedge I_{fh} = 1\}$ linked with f are included in the household sample.

If the household h is included in S_H^3, then the set of farms $S_F^3 = \{f \in F \mid h \in S_H^3 \wedge I_{fh} = 1\}$ linked with h are included in the farm sample. If f is in $\{S_F^q; q = 1,2,3\}$, then the set of parcels $L_f = \{l \in L \mid I_{fl} = 1\}$ are included in the sample. This sampling procedure is related with both, network sampling and indirect sampling [Falorsi (2014), Singh and Mecatti (2011), Mecatti and Singh (2014)].

The parameter to be estimated is the population total, $P = \{L, F, H\}$: over land, $Y_L = \sum_{l=1}^{L} Y_l$, over farms, $Y_F = \sum_{f=1}^{F} Y_f$, and over households, $Y_H = \sum_{h=1}^{H} Y_h$. Given the links (I_{lf}, I_{fh}) between (l, f, h), (i) the multiplicity of the parcel l is $m_l = \sum_{q=1}^{Q} m_l^q$, where $m_l^1 = \sum_{i=1}^{A^1} I_{il}^1$, $m_l^2 = \sum_{i=1}^{A^2} I_{if}^2 I_{lf} = m_f^2 I_{lf}$ and $m_l^3 = \left(\sum_{h=1}^{H} m_h^3 I_{fh}\right) I_{lf}$; (ii) the multiplicity of the farm f is $m_f = \sum_{q=1}^{Q} m_f^q$, where $m_f^1 = \sum_{i=1}^{A^1} I_{if}^1$, $m_f^2 = \sum_{i=1}^{A^2} I_{if}^2$ and $m_f^3 = \sum_{h=1}^{H} m_h^3 I_{fh}$; and the multiplicity of the household h is $m_h = \sum_{q=1}^{Q} m_h^q$, where $m_h^1 = \sum_{f=1}^{F} m_f^1 I_{fh}$, $m_h^2 = \sum_{f=1}^{F} m_f^2 I_{fh}$ and $m_h^3 = \sum_{i=1}^{A^3} I_{ih}^3$.

The total over land is $Y_L = \sum_{q=1}^{Q} \tilde{Y}_{Lq}$, where $\tilde{Y}_{Lq} = \sum_{i=1}^{A^q} \tilde{y}_{Li}^q$, where $\tilde{y}_{Li}^1 = \sum_{l=1}^{L} I_{il}^1 \frac{y_l}{m_l}$, $\tilde{y}_{Li}^2 = \sum_{f=1}^{F} I_{if}^2 \sum_{l=1}^{L} I_{lf} \frac{y_l}{m_l}$ and $\tilde{y}_{Li}^3 = \sum_{f=1}^{F} \left(\sum_{h=1}^{H} I_{ih}^3 I_{fh}\right) \sum_{l=1}^{L} I_{lf} \frac{y_l}{m_l}$ are the multiplicity-adjusted values of the survey variable associated to the i^{th} sampling unit in each frame. The total over farms is $Y_F = \sum_{q=1}^{Q} \tilde{Y}_{Fq}$, where $\tilde{Y}_{Fq} = \sum_{i=1}^{A^q} \tilde{y}_{Fi}^q$, where $\tilde{y}_{Fi}^1 = \sum_{f=1}^{F} I_{if}^1 \frac{y_f}{m_f}$, $\tilde{y}_{Fi}^2 = \sum_{f=1}^{F} I_{if}^2 \frac{y_f}{m_f}$ and $\tilde{y}_{Fi}^3 = \sum_{f=1}^{F} \left(\sum_{h=1}^{H} I_{ih}^3 I_{fh}\right) \frac{y_f}{m_f}$. The total over households is $Y_H = \sum_{q=1}^{Q} \tilde{Y}_{Hq}$, where $\tilde{Y}_{Hq} = \sum_{i=1}^{A^q} \tilde{y}_{Hi}^q$ and $\tilde{y}_{Hi}^1 = \sum_{f=1}^{F} I_{if}^1 \sum_{h=1}^{H} I_{fh} \frac{y_h}{m_h}$, $\tilde{y}_{Hi}^2 = \sum_{f=1}^{F} I_{if}^2 \sum_{h=1}^{H} I_{fh} \frac{y_h}{m_h}$, and $\tilde{y}_{Hi}^3 = \sum_{h=1}^{H} I_{ih}^3 \frac{y_h}{m_h}$.

The multiplicity-adjusted estimator, $\hat{Y}_P = \sum_{q=1}^{Q} \sum_{i=1}^{S^q} w_i^q \tilde{y}_{Pi}^q$, is unbiased and its variance is $V\hat{Y}_P = \sum_{q=1}^{Q} \sum_{i=1}^{A^q} \sum_{i'=1}^{A^q} \left(\pi_{ii'}^q - \pi_i^q \pi_{i'}^q\right) \frac{\tilde{y}_{Pi}^q}{\pi_i^q} \frac{\tilde{y}_{Pi'}^q}{\pi_{i'}^q}$. The variance estimator is $\hat{V}\hat{Y}_P = \sum_{q=1}^{Q} \sum_{i=1}^{S^q} \sum_{i'=1}^{S^q} \frac{\pi_{ii'}^q - \pi_i^q \pi_{i'}^q}{\pi_{ii'}^q} \frac{\tilde{y}_{Pi}^q}{\pi_i^q} \frac{\tilde{y}_{Pi'}^q}{\pi_{i'}^q}$.

The multiplicity-adjusted estimator can be written in terms of population units as an adjusted-weight estimator, $\hat{Y}_{\mathrm{P}} = \sum_{q=1}^{Q}\sum_{j=1}^{S_{\mathrm{P}}^q} \tilde{w}_j^q y_j$, where $\tilde{w}_j^q = \frac{1}{m_j}\sum_{i=1}^{S^q} w_i^q$.

4. Multiple frame regression estimators

To use auxiliary information, we specify a regression model in terms of population units, $y_j = \mathbf{x}_j\boldsymbol{\beta}+\varepsilon_j$, where $\mathbf{x}_j$ is the $(1\times p)$ vector of auxiliary variables, including the constant 1, $\boldsymbol{\beta}$ is a $(p\times 1)$ vector of regression parameters, $E\varepsilon_j = 0$, and $V\varepsilon_j = \sigma^2$. The model in terms of sampling units is, $\tilde{y}_i^q = \tilde{\mathbf{x}}_i^q\boldsymbol{\beta} + \tilde{\varepsilon}_i^q$, where $\tilde{\mathbf{x}}_i^q = \sum_{j=1}^{S_{\mathrm{P}}^q}\alpha_{ij}^q\mathbf{x}_j$, $\tilde{\varepsilon}_i^q = \sum_{j=1}^{S_{\mathrm{P}}^q}\alpha_{ij}^q\varepsilon_j$, $E\tilde{\varepsilon}_i^q = 0$, $V\tilde{\varepsilon}_i^q = \sigma^2\sum_{j=1}^{S_{\mathrm{P}}^q}\left(\alpha_{ij}^q\right)^2$.
Lu (2014) proposes four methods to estimate $\boldsymbol{\beta}$. We consider the probability weighted least square estimator, $\hat{\boldsymbol{\beta}}_{\tilde{w}} = \min_{\boldsymbol{\beta}}\sum_{q=1}^{Q}\sum_{i=1}^{S^q}\tilde{w}_i^q\left(\tilde{y}_i^q - \tilde{\mathbf{x}}_i^q\boldsymbol{\beta}\right)^2$, where $\tilde{w}_i^q = w_i^q\tilde{\alpha}_i^q$ and $\tilde{\alpha}_i^q = \frac{1}{\sum_{j=1}^{S_P^q}\left(\alpha_{ij}^q\right)^2}$: it is

$\hat{\boldsymbol{\beta}}_{\tilde{w}} = \left(\tilde{\mathbf{X}}^{\mathbf{T}}\mathbf{D}_{\tilde{w}}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^{\mathbf{T}}\mathbf{D}_{\tilde{w}}\tilde{\mathbf{y}}$, where $\tilde{\mathbf{X}}$ is the $\left(\sum_{q=1}^{Q}S^q\times p\right)$ multiplicity-adjusted auxiliary data matrix,

$\tilde{\mathbf{y}}$ is the $\left(\sum_{q=1}^{Q}S^q\times 1\right)$ vector of multiplicity-adjusted survey variable data,

and $\mathbf{D}_{\tilde{w}} = diag\left\{\tilde{w}_i^q; i=1,2,\cdots,S^q; q=1,2,\cdots,Q\right\}$.

$\hat{\boldsymbol{\beta}}_{\tilde{w}}$ is a design-consistent estimator of the regression parameter values in the finite population, $\boldsymbol{\beta}_N = \left(\tilde{\mathbf{X}}_N^T\tilde{\mathbf{X}}_N\right)^{-1}\tilde{\mathbf{X}}_N^T\tilde{\mathbf{y}}_N$, where $N = \sum_{q=1}^{Q}A^q$ is the number of sampling units in the multiple frame, $\tilde{\mathbf{X}}_N$ is the $(N\times p)$ matrix of multiplicity-adjusted auxiliary variable values, and $\tilde{\mathbf{y}}_N$ is the $(N\times 1)$ vector of the multiplicity-adjusted survey variable values.

The Multiplicity-adjusted Generalized Regression estimator (MGREG) is $\hat{Y}_{MGREG} = \sum_{q=1}^{Q}\sum_{i=1}^{A^q}\tilde{\mathbf{x}}_i^q\hat{\boldsymbol{\beta}}_{\tilde{w}}$: it is a design-consistent estimator of the population total, Y , and its asymptotic design-variance can be estimated using $\hat{V}\hat{Y}_{MGREG} = \hat{V}\left(\sum_{q=1}^{Q}\sum_{i=1}^{S^q}w_i^q\hat{e}_i^q\right)$g, where $\hat{e}_i^q = \tilde{y}_i^q - \tilde{\mathbf{x}}_i^q\hat{\boldsymbol{\beta}}_{\tilde{w}}$ (Fuller, 2009; Kim and Rao, 2012).
Ranalli et al (2014) propose calibration estimators. Deville and Särdal (1992) and (Fuller, 2009) show how calibration estimators can be approximated by regression estimators.

Integrating survey and register data

The MGREG estimator is useful to integrate survey and register data. To see this, we assume that there is a set of values $\left(y_j, \mathbf{x}_j\right)$ associated with each population unit: y_j is the survey variable value and $\mathbf{x}_j$ are register values. We assume that the choice of $\mathbf{x}_j^q$ differs among single frames (registers) and we use a different working model in each register, $\tilde{y}_i^q = \tilde{\mathbf{x}}_i^q \boldsymbol{\beta}^q + \tilde{\varepsilon}_i^q$, where $\tilde{\mathbf{x}}_i^q = \sum_{j=1}^{S_P^q} \alpha_{ij}^q \mathbf{x}_j^q$, $\tilde{\varepsilon}_i^q = \sum_{j=1}^{S_P^q} \alpha_{ij}^q \varepsilon_j$, $E\tilde{\varepsilon}_i^q = 0$, $V\tilde{\varepsilon}_i^q = \sigma^{2,q} \sum_{j=1}^{S_P^q} \left(\alpha_{ij}^q\right)^2$. To observe data on $\left(y_j, \mathbf{x}_j^q\right)$, we consider Q^1 frames of the target population, P, and we select independently from each one a sample, $\left\{S^q; q = 1, 2, \cdots, Q^1\right\}$. We consider Q^2 registers as independent large samples, $\left\{S^q; q = 1, 2, \cdots, Q^2\right\}$, selected from P, where we observe only data on $\mathbf{x}_j^q$.

To estimate regression parameters, $\boldsymbol{\beta}^q$, we use data from Q^1 and the probability weighted least square estimator, $\hat{\boldsymbol{\beta}}_{\tilde{w}}^q = \min_{\boldsymbol{\beta}^q} \sum_{i=1}^{S^q} \tilde{w}_i^q \left(\tilde{y}_i^q - \tilde{\mathbf{x}}_i^q \boldsymbol{\beta}^q\right)^2$, which is $\hat{\boldsymbol{\beta}}_{\tilde{w}}^q = \left(\tilde{\mathbf{X}}^{qT} \mathbf{D}_{\tilde{w}}^q \tilde{\mathbf{X}}^q\right)^{-1} \tilde{\mathbf{X}}^{qT} \mathbf{D}_{\tilde{w}}^q \tilde{\mathbf{y}}^q$, where $\tilde{\mathbf{X}}^q$ is the $\left(S^q \times p^q\right)$ multiplicity-adjusted auxiliary data matrix, $\tilde{\mathbf{y}}^q$ is the $\left(S^q \times 1\right)$ vector of multiplicity-adjusted survey variable data, and $\mathbf{D}_{\tilde{w}}^q = diag\left\{\tilde{w}_i^q; i = 1, 2, \cdots, S^q\right\}$.

We use data from Q^2 to estimate $\sum_{i=1}^{A^q} \tilde{\mathbf{x}}_i^q$, using $\sum_{i=1}^{S^q} w_i^q \tilde{\mathbf{x}}_i^q$. The MGREG estimator is $\hat{\hat{Y}}_{MGREG} = \sum_{q=1}^{Q^2} \sum_{i=1}^{S^q} w_i^q \tilde{\mathbf{x}}_i^q \hat{\boldsymbol{\beta}}_{\tilde{w}}^q$, and its error is

$$\hat{\hat{Y}}_{MGREG} - Y = \hat{\hat{Y}}_{MGREG} - \sum_{q=1}^{Q^2} \tilde{\mathbf{x}}_{N^q}^q \boldsymbol{\beta}_{N^q}^q = \sum_{q=1}^{Q^1} \sum_{i=1}^{S^q} \left(\tilde{y}_i^q - \tilde{\mathbf{x}}_i^q \boldsymbol{\beta}_{A^q}^q\right) + \sum_{q=1}^{Q^2} \left(\sum_{i=1}^{S^q} w_i^q \tilde{\mathbf{x}}_i^q - \tilde{\mathbf{x}}_{A^q}^q\right) \hat{\boldsymbol{\beta}}_{\tilde{w}}^q + \left(\tilde{\mathbf{x}}_N - \sum_{q=1}^{Q^1} \sum_{i=1}^{S^q} w_i^q \tilde{\mathbf{x}}_i^q\right) \left(\hat{\boldsymbol{\beta}}_{\tilde{w}}^q - \boldsymbol{\beta}_{A^q}^q\right)$$

, where $\tilde{\mathbf{x}}_{A^q}^q = \sum_{i=1}^{A^q} \tilde{\mathbf{x}}_i^q$, $\tilde{\mathbf{x}}_N = \sum_{q=1}^{Q^2} \tilde{\mathbf{x}}_{A^q}^q$, $\boldsymbol{\beta}_{A^q}^q = \left(\tilde{\mathbf{X}}_{A^q}^{qT} \tilde{\mathbf{X}}_{A^q}^q\right)^{-1} \tilde{\mathbf{X}}_{A^q}^{qT} \tilde{\mathbf{y}}_{A^q}^q$, $\tilde{\mathbf{X}}_{N^q}^q$ is the $\left(A^q \times p^q\right)$ matrix of multiplicity-adjusted auxiliary variable values, and $\tilde{\mathbf{y}}_{A^q}$ is the $\left(A^q \times 1\right)$ vector of the multiplicity-adjusted survey variable values.

$\hat{\hat{Y}}_{MGREG}$ is design-consistent and its asymptotic design-variance can be estimated using $\hat{V}\hat{\hat{Y}}_{MGREG} = \hat{V} \sum_{q=1}^{Q^1} \sum_{i=1}^{S^q} w_i^q \hat{\tilde{e}}_i^q + \sum_{q=1}^{Q^2} \hat{\boldsymbol{\beta}}_{\tilde{w}}^{qT} \left(\hat{\mathbf{V}} \sum_{i=1}^{S^q} w_i^q \tilde{\mathbf{x}}_i^q\right) \hat{\boldsymbol{\beta}}_{\tilde{w}}^q$, where $\hat{\tilde{e}}_i^q = \tilde{y}_i^q - \tilde{\mathbf{x}}_i^q \hat{\boldsymbol{\beta}}_{\tilde{w}}^q$. The elements of the covariance matrix, $\hat{\mathbf{V}} \sum_{i=1}^{S^q} w_i^q \tilde{\mathbf{x}}_i^q$, can be estimated using the HT variance estimator. If $A^q \in Q^2$ is complete, then $\sum_{i=1}^{A^q} \tilde{\mathbf{x}}_i^q$ is known and all terms in the covariance matrix related with A^q are nulls.

Analysis of complex surveys: sampling design informativeness

Linear (regression) and generalized linear models are useful tools for analyzing survey data. Deaton (1997) shows how they can be used with household surveys and with linked farm-household surveys (Singh et al., 1986). Most land use models are generalized linear models (Ambrosio et al., 2008), useful for analysing linked farm-parcel surveys. Relative little work has been done on 'sustainometrics' models (Todorov and Marinova, 2010), for analysing linked farm-household-parcel surveys.

Typically, the analysts fit these models assuming that the sampling design is 'non informative'. However, complex sampling design leads usually to informative samples and, as a result, model parameters estimator are inconsistent (Binder et al, 2005). The weighted estimator is consistent and its asymptotic distribution is normal, and can be used for hypothesis testing and prediction [Fuller (2009)].

5. Concluding remarks

To collect data for designing sustainable policies, we integrate agricultural and household sampling frames in a multiple overlapping frame. This sampling frame provides the required reporting units: the farm as economic unit, the household as social unit, and the parcel as environmental unit. We apply this strategy in three Latin America countries.

Due to overlapping, the sum of the usual single frame weight estimator over the multiple frames is biased. The multiple frame estimators proposed in the literature consist in an adjustment of the single frame weight estimator by adjusting either the sampling weight (adjusted-weight estimators) or the survey variable (multiplicity-adjusted estimators). We focus on multiplicity-adjusted estimators because they are internally consistent and take into account the linkage among sampling units in a simple way.

Regression estimators have received considerable attention in the literature, but mainly for single frames. We extend this single frame estimator to the multiple frame case, highlighting the usefulness of multiple frame regression estimators to integrate register and survey data.

Linear (regression) and generalized linear models are useful tools for analysing survey data, and we suggest using weighted estimators to estimate model parameters, in order to take into account the sampling design informativeness. Hypothesis testing and prediction can be carried out using the asymptotic normal distribution of these estimators.

6. REFERENCES

Ambrosio L. Iglesias L., Marín C., Pascual V., and Serrano A. (2008). A spatial high-resolution
model of agricultural land use dynamics. *Agricultural Economics*, **38**:233-45.

Ambrosio L. (2013): Marco de muestreo y diseño de la Encuesta Nacional Agropecuaria de
Guatemala. Informe Técnico. FAO. Universidad Politécnica de Madrid.

Ambrosio L. (2014): Diagnóstico del actual sistema de estadísticas agropecuarias y marco conceptual y metodológico para estadísticas agropecuarias en Ecuador. Informe Técnico. FAO. Universidad Politécnica de Madrid.

Ambrosio L. and Iglesias L. (2014) Identifying the most appropriate sampling frame for specific landscape types. Technical Report Series. GO-01-2014. FAO.

Ambrosio L. (2015): Marco de muestreo y muestra maestra para encuestas integradas y vinculadas en Costa Rica. Informe Técnico. FAO. Universidad Politécnica de Madrid.

Bankier, M.D. 1986. Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys. *Journal of the American Statistical Association*, **81**: 1074-1079.

Binder, D.A., Kovacevic, M.S. and Roberts G. (2005). How important is the informativeness of the sampling design. *Proceedings of the Survey Methods Section*, pp 1-11.

Deaton, A. (1997). The analysis of household surveys. A microeconometric approach to development policy. World Bank. Johns Hopkins University Press

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**: 376-382.

FAO (1996). Multiple frame agricultural surveys. Vol.1. Current surveys based on area and list sampling methods. Statistical Development Series. 7. Rome.

FAO (1998). Multiple frame agricultural surveys. Vol.2. Agricultural survey programmes based on area frame or dual frame (area and list) sample designs. Statistical Development Series. 10. Rome.

FAO. World Bank and United Nations Statistical Commission (2011). Global Strategy to Improve Agricultural and Rural Statistics. The World Bank.

FAO. World Bank and United Nations Statistical Commission (2012). Action Plan of the Global Strategy to Improve Agricultural and Rural Statistics. FAO. Rome.

FAO (2015). Handbook on master sampling frame for agriculture. Technical Report Series. GO-01-2015.

Falorsi P.D. (2014) Integrated survey framework. Technical Report Series GO-02-21014. FAO Statistics Division. Rome

Fuller,W.A. (2009). Sampling statistics. Wiley. New York.

Faulkenberry, G.D., Garoui, A. (1991): Estimating a population total using an area frame. *Journal of the American Statistical Association*, **86** : 445-449.

Fecso R., Tortora R. D. and Vogel F. (1986). Sampling Frames for Agriculture in the United States. *Journal of Official Statistics*, **2:**279-292.

Hartley, H. O. (1962). Multiple Frame Surveys. *Proceedings of the Social Statistics Section*. American Statistical Association.

Hartley, H. O. (1974). Multiple Frame Methodology and Selected Applications. *Sankhya*, Ser. C, **36**: 99-118.

Kalton G. and Anderson D.W. (1986) Sampling rare populations, *Journal of the Royal statistical Society*, Series A, **149**: 65-82

Kim J.K. and Rao J.N.K. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*, **99**: 85-100

Lohr, S. and Rao, J. N. K. (2000). Inference from Dual Frame Surveys. *Journal of the American Statistical Association*. **95**: 271-280.

Lohr, S. and Rao, J. N. K. (2006). Estimation in Multiple-Frame Surveys. *Journal of the American Statistical Association*. **101**: 1019-1030

Lohr S. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology*, **37**: 197-213

Lu Y. (2014). Regression coefficient estimation in dual frame surveys. *Communication in Statistics-Simulation and Computation*, **43**: 1675-84

Mecatti F. and Singh A.C. (2014). Estimation in multiple frame surveys: A simplified and unified review using multiplicity approach. *Journal de la Société Française de Statistique*, **155**: 51-69

Ranalli, M. G., Arcos, A., Rueda, M. d. M., and Teodoro, A. (2014). Calibration estimation in dual frame surveys. arXiv preprint arXiv:1312.0761v2.

Singh J, Squiere L., and Strauss J (1986) Agricultural household models. World Bank.

Singh A. and Mecatti F. (2011). Generalized multiplicity-adjusted Horvitz-Thompson type estimation as a unified approach to multiple frame survey. *Journal of Official Statistics*, **27**: 633-650

Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, **91**: 349-356.

Todorov V. and Marinova D. (2011). Modelling sustainability. *Mathematics and Computers in Simulation*, **81**: 1397-1408.

UNSD (1986). National Household Survey Capability Program. Sampling Frames and Sample Designs for Integrated Household Survey Programs. Department of Technical Co-Operation for Development and Statistical Office. United Nations. New York.

UNSD (2008). Designing Household Survey Samples: Practical Guides. ST/ESA/STAT/SER.F/98 Department of Economic and Social Affairs. Statistics Division. Studies in Methods. Series F Nº 98. Unite ations. New York

Vogel F.A. (1995). The evolution and development of agricultural statistics at the United States Department of Agriculture. *Journal of Official Statistics*, **11:**161-180.