Improving methods for linking area frames with list frames: literature review preliminary results

Cristiano Ferraz
Federal University of Pernambuco, Brazil
cferraz@de.ufpe.br

Abstract

This paper is part of the activities related to FAO's Global Strategy Research Plan for 2013, and describes the preliminininary results of a literature review concerning methods for linking area frames with list frames, focusing on the production of agricultural and rural statistics.

1 Introduction

Developed as part of the FAO's 2013 Global Strategy Research Plan activities, this paper introduces a literature review on methods for combining information from area frames and list frames with applications to agricultural surveys and censuses.

Agricultural surveys and censuses are known to use area frames, list frames, and also both in a dual frame design to support the efficient production of official statistics. In a general scenario, efforts to take advantage of relevant information raised by linking data from different types of frames can lead to improvements in the agriculture and rural statistics precision. Combining information from different frame sources also contributes to the identification and building of an appropriate master frame for integrated agricultural surveys. Revisiting papers and technical reports related to the subject, we expect to identify the main advantages and disadvantages of various types of list and area frames, as well as the advantages and disadvantages of several estimators found in the literature. The goal is to provide guidance on the most appropriate type of frame and estimation process to produce agricultural statistics.

This paper is divided in five sections including the introduction. In Section 2, basic concepts related to area and list frames are introduced. The potential role of administrative data for generating and improving frames is also considered. In Section 3, the subject of combining information from area frames and list frames by dual frame designs is introduced. A literature review on point estimators is presented within the context of producing agricultural statistics. Section 4 briefly introduces the record linkage subject using a dual frame scenario as motivation. Section 5 presents conclusion remarks, recommendations, and outlines areas where further research is needed.

2 Area Frames, list frames and administrative registers

In this section, a brief review of main concepts related to area and list frames is provided, and the subject of using administrative records to produce agricultural statistics is briefly discussed.

Frame is a fundamental concept in survey sampling. Probability sampling procedures are applied to the frame component units that can be elements or sets of elements (clusters) from the target population. The choice and adoption of a good sampling design depend on the nature of the component units, as well as on the nature of any eventual auxiliary information contained in the available frame. The effectiveness of the chosen sampling design, in the

sense of providing accurate and precise estimates of the parameters of interest, depend on the quality of the frame in use. In a perfect situation, the frame is supposed to provide full coverage and unique identification of all the target population elements. In practice, statisticians have to deal with less than perfect situations, as several limitations can affect the quality of the frames. Ignoring or failing to properly identifying these limitations can lead to biased estimates, inflated variances, or even both.

In cases where the frame component units are segments of land areas, the frame is called an area frame, and the sampling procedure applied to it is generally called area sampling. When information about the area measure of each segment is available, sampling with probability proportional to the area measure is a common choice of sampling design applied to area frames. In several circunstances, on the other hand, the frame components are simply lists of addresses for holders or farmers, justifying the expression list frames. Considering only the type of frame used, agricultural survey designs can be classified into two basic categories: area sampling designs and list sampling designs.

Adequate frames can be build based on information collected in agricultural and population censuses carried out successfully. Maintaining their adequacy along time, however, depend on the availability of an efficient information system that could proceed updating the data. With constant advances in information technology, the role of administrative registers in feeding such systems can be feasibly explored, and even the possibility of using surveys based only on administrative data to generate agricultural statistics, evaluated. Selander et al. [18] and Wallgren and Wallgren [20], investigated the possibility of producing statistics on crops and livestock using the ICAS - Integrated Administrative and Control System, an European system of information relate to agricultural subsidies. Carfagna and Carfagna [5] analyzed the advantages and disadvantages of producing agricultural statistics based on administrative data, comparing with survey samples. Designing a survey based only on administrative data differs from designing surveys based on sample data in many ways. While in sample surveys the data collection process happens after defining aspects of sampling design such as target population definition and parameters of interest, in administrative data surveys the data collection process preceeds such aspects, as data has already been collected. Although the use of administrative data to produce agricultural statistics brings a series of chalenges concerning proper coverage, managing data quality and identifying adequate variables of interest, their joint use with survey data can be very promising to improve agricultural surveys design, as shown by Lavalle [15]. Identifying methods to improve the matching of several sources of data at the frame components unit level can help the feasibility of good quality frames, not only in terms of coverage rates but also in terms of providing access to auxiliary information.

2.1 Sampling frames for agricultural surveys

Agricultural surveys based on non-probability sampling designs are commonly called subjective sampling surveys. We focus on the description of the main characteristics of frames for agricultural probability sample surveys.

2.1.1 List frames

Usually formed by a list of holdings or holders addresses, list frames are the most common type of frame found in agricultural probability sample surveys. Sometimes their component units are clusters, and multiple stage sampling design is necessary to select elements (holders). They can be build upon information collected from the most recent agricultural or population census, administrative data, previous surveys or a combination of several data sources.

Often, auxiliary information is available in list frames, allowing the use of efficient sampling schemes such as stratified sampling or probability proportional to size sampling. However, list frames are subject to great degeneration over time, leading to serious coverage and out of date auxiliary information problems if they are not properly maintained. Problems with duplicated records should also be avoided.

Countries with advanced information systems, based on administrative data can take advantage of combinig information from various sources to build efficient list frames. The high cost of the building process, however, can be an issue. Althoug studies to estimate its magnitude are important, carrying out them is a challenge. D'Orazio et al. [6] review a

series of methods aiming to match records between files. The potential impact of matching processes errors on statistical estimates is a subject for which studies should be conducted.

2.1.2 Area frames

The segments of area that compose area frames can be determined in three different ways. They can be set by identifiable physical boundaries such as rivers, roads etc., by a squared grid of map coordinates or by coinciding their limits with the ones of agricultural holding lands (FAO Statistical Development Series 7 [8], page 20). If the segment do not coincide with the boundaries of a holding, it is necessary to define a tract. Segments are subdivided into non-overlapping tracts. One tract is the part of a holding that is found within the limits of a segment, or a peice of land that do not belong to any holding. A holding comprises at least one tract. Tracts are observational units.

Area frames have the advantage of providing full coverage of the target population, are duplication free and remain updated for a long time. However, they are expensive to sample from, and sensible to the impact of outliers. Since very often stratification is used in area frames, refreshing information for such task requires efforts on raising updataed relevant auxiliary information. The possibility of using administrative record to raise the necessary data should be evaluated carefuly. Carfagna [4] describes the results of the 2001 AGRIT project, in which the efforts to change from an area frame with clusters to an area frame with unclustered points, by using ICAS data to refresh stratfication variables were assessed. The 1.2 million points of the new sampling frame were stratified by the classes: arable land, permanent crops, permanent grass, forest, isolated trees and rural buildings, other (artificial areas, water, etc.). The study concluded that it would not be worth to annualy proceed with the necessary updating procedure to change the area frame feature, as there was low accuracy on the land classification assessed by means of a sample.

When efforts to build a list frame based on different sources of data are not sufficient to guarantee full population coverage, an area frame should be adopted in order to avoid bias. Area frame is also a great choice for multipurpose studies in which parameters of interest are related to land areas, such as it happens with agriculture and environmental surveys. Regarding agricultural surveys, an important feature of area frames is that they allow the assessment of characteristics that can be observed on the ground, such as cultivated area.

Although expensive, the materials necessary to use in surveys with area frames are becoming cheaper and more accurate.

A summary of the types of area and list frame suitable for agricultural surveys is provided in Table 1.

Table 1: types of area and list frame suitable for agricultural surveys

Frame type	Frame description	Unit component	Unit type example
1	List frame	Element	Holder addresses
2	List frame	Cluster	Villages
3	Area frame	Segment (element)	Holding area
4	Area frame	Map grid (cluster)	Point
5	Area frame	Land area (cluster)	Physical boundaries

3 Combining list and area frames information through dual frame designs

In agriculture, and several other types of surveys, area frames and list frames can be simultaneously used to select a sample using a dual frame approach. Succesful examples in ecology and health, for instance, are described by Haines and Pollock [12], and Shin et al. [17]. In agriculture, examples of programmes that used the two frames combined are described at FAO's Statistical Development Series 10 [9].

In general, dual frame surveys use two overlapping frames to cover the same target population. Independent samples are selected from each one, allowing different probability sampling schemes applied in each frame. It is common to find agriculture applications where

a stratified area sampling is selected from an area frame, and a random sample is drawn from a list frame of farm operators. In this section, several scenarios of dual frame designs for agricultural surveys are analyzed, examining their main requirements for implementation, advantages and disadvantages.

3.1 General design characteristics

Several dual frame estimators can be found in the literature. Their differences are essentialy in the way they take advantage of information coming from both frames. In this sense, dual frame inference may be viewed as a form of linking two frames in the estimation stage of a survey.

Let the target population U be covered by two frames $\mathcal{A}$ and $\mathcal{B}$. Inference in dual frame surveys, as originaly proposed by Hartley [13], considers the broad scenario illustrated by Figure 1, where three domains can be identified: $a = \mathcal{A} \cap \mathcal{B}^c$, $b = \mathcal{B} \cap \mathcal{A}^c$ and $ab = \mathcal{A} \cap \mathcal{B}$. Two importante requirements are needed:

- Completeness; and
- Identifiability.

The completeness requirement is met when both frames used in the design provide full coverage to the target population, i.e. $U = \mathcal{A} \cup \mathcal{B}$. In this way, every element is listed in at least one of the frames. The identifiability requirement is achieved when, for any sampled element, it is possible to tell if it belongs or not to each frame. Although the dual frame idea can be generalized to multiple frames, it is not difficult to see that dealing with more than two frames increases the challenge of meeting the necessary requirements.

Dual frame surveys have the flexibility to accomodate different types of frames. Figure 1 could represent the use of two incomplete list frames. On the other hand, if frame $\mathcal{A}$ is an area frame, and frame $\mathcal{B}$ is a list frame, for example, then a particular case where $\mathcal{B} \subset \mathcal{A}$ happens, as illustrated by Figure 2. Using a dual frame approach in such cases can save money if the costs of sampling from frame B are cheaper than frame A, or improve accuracy if auxiliary information is available from frame $\mathcal{B}$.

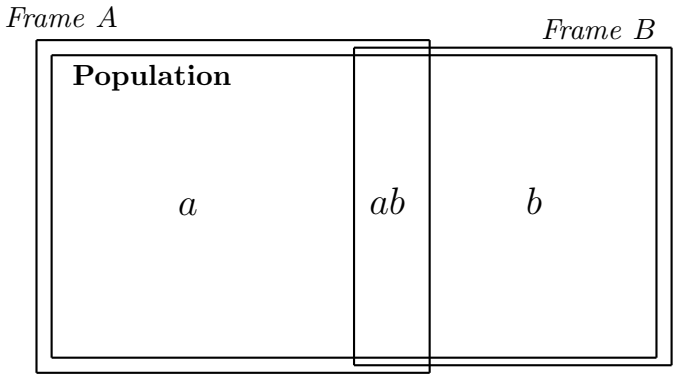

Figure 1: General dual frame scenario

The choice of an appropriate sampling design to each frame depend on the amount of information available and the capacity of identifying the elements in the domains. If efficient linkage methods can be applied to the frames for matching their records and identifying the elements in each domain, stratification of the population is feasible, and probability samples of fixed sizes n_a, n_b and n_{ab} can be selected from the domains indicated by their respective indices. This situation corresponds to building a unique frame, unifying information from frames A and B. If the record linkage is not an option, probability samples of sizes n_A and n_B are selected from each frame. In any case, estimators for t, a population total of interest, can be written as the sum of estimators for each domain total:

$$\hat{t} = \hat{t}_a + \hat{t}_b + \hat{t}_{ab}, \tag{1}$$

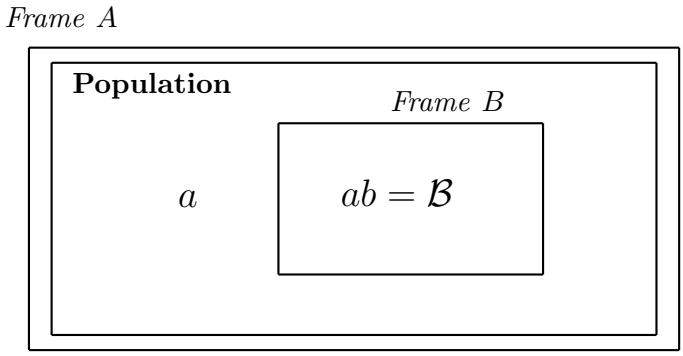

Figure 2: Special case dual frame scenario based on area frame and list frame

where $\hat{t}$ denotes a dual frame estimator for the population total, and $\hat{t}_a$, $\hat{t}_b$, and $\hat{t}_{ab}$ denote estimators for the domains totals. Variances forms depend on how each estimator deal with information raised from both frames. In general,

$$Var\left(\hat{t}\right) = Var(\hat{t}_a) + Var(\hat{t}_b) + Var(\hat{t}_{ab}) + Cov(\hat{t}_a, \hat{t}_{ab}) + Cov(\hat{t}_b, \hat{t}_{ab}) \tag{2}$$

In the case where application of record linkage methods allows the population to be stratified, the covariance terms are null. The simply possibility of matching the records from both frames, however, does not necessarily indicate that doing so is the best decision. The complexity of the matching task and all the potential problems related with the operation should be kept in mind.

For complex sampling designs applied to the frames, jackknife estimation of variances can be used for Hartley's and the next dual frame estimators, as considered by Lohr and Rao [16]. Taylor linearization thecniques are also an option. Approximated variances of dual frame estimator's written as in 1 can be presented in the general matricial form

$$A\text{Var}\left(\hat{t}\right) = \mathbf{d}^T \mathbf{\Sigma} \mathbf{d}, \tag{3}$$

where $\mathbf{d}^T$ is a column vector, and $\mathbf{\Sigma}$ is an appropriate block-diagonal covariance matrix.

Carfagna [3] provides a broad review of the main advantages, disadvantages and requirements of dual frame designs. In this paper, we focus on a review of four major estimators proposed at the literature, and outline their potential use according to the types of agricultural frames described in Table 1 available for the survey.

3.2 Dual frame estimators review

Let π_k^A and π_k^B be the first order inclusion probabilities for elements of each frame in a dual frame survey, and let y_k be the value of the variable of interest for $k \in U$.

Let also U_{ab} denote the set of population elements belonging to domain ab, while S_{ab}^A denote the sample set of elements from U_{ab} selected from frame A. Whenever necessary, superscripts A or B indicate the frame from which information is related to. For instance, $\overline{y}_{ab}^B$ denote the domain ab sample mean from frame B. Some dual frame estimators found in the literature are introduced next.

3.2.1 Hartley's estimator

The estimator proposed by Hartley [13] can be expressed as a weighted average between appropriate Horvitz-Thompson [14] estimators applied to each dual frame domain:

$$\hat{t}_H = \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \sum_{k \in S_b} \frac{y_k}{\pi_k^B} + \sum_{k \in S_{ab}} y_k^*. \tag{4}$$

In this expression,

$$y_k^* = \begin{cases} p\dfrac{y_k}{\pi_k^A}, & \text{if } k \in S_{ab}^A \\ (1-p)\dfrac{y_k}{\pi_k^B}, & \text{if } k \in S_{ab}^B \end{cases} \tag{5}$$

and p is a weighting constant ($0 \leq p \leq 1$) chosen to minimize the variance of the estimator $\hat{t}_H$. In order to applied the estimator, the requirement of identifiability is crucial.

Considering an agricultural survey with frame A as an area frame, and frame B as a list frame, Hartley's estimator can be immediately applied in surveys where frame A is of type 3 and frame B is of type 1. In this situation, elements (holder's addresses, for instance) can be directly selected using one-stage probability sample designs such as simple random sampling. If at least one of the frames has clusters as elements (types 2, 4 or 5), multiple-stage sampling design are necessary to achieve the elements, and then make the matching of sampled units feasible (identifieability requirement).

One advantage of estimator 4 is the possibility of simplifying the estimation process by chosing the value of p to be zero. In such case, the resulting estimator is called a screening estimator because in order to be effective, the procedure requires to screen and eliminate from frame A, every component unit that also belongs to frame B.

Among the disadvantages of the estimator proposed by Hartley are the fact that, for direct element sampling designs, information about N_a, N_b and N_{ab}, the population domain sizes, are assumed available. Supposing simple random sampling is applied to each frame, and denoting the domain population variances by σ_a^2, σ_b^2 and σ_{ab}^2, the variance of 4 is approximated by the variance for stratified samples with allocation proportional to the domain sizes (ignoring finite population correction factors):

$$Var(\hat{t}_H) = \frac{N_A^2}{n_A}\left[\sigma_a^2\left(1-\frac{N_{ab}}{N_A}\right) + p^2\sigma_{ab}^2\frac{N_{ab}}{N_A}\right] + \frac{N_B^2}{n_B}\left[\sigma_b^2(1-\frac{N_{ab}}{N_B}) + (1-p)^2\sigma_{ab}^2\frac{N_{ab}}{N_B}\right]. \tag{6}$$

Using complex sample designs, the approximated variance 2 can be applied. The approximated variance components of Hartley's estimator are:

$$\mathbf{d} = \begin{bmatrix} 1 \\ p \\ 1 \\ (1-p) \end{bmatrix} \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_a^2 & \sigma_{a;ab(A)} & 0 & 0 \\ \sigma_{a;ab(A)} & \sigma_{ab(A)}^2 & 0 & 0 \\ 0 & 0 & \sigma_b^2 & \sigma_{b;ab(B)} \\ 0 & 0 & \sigma_{b;ab(B)} & \sigma_{ab(B)}^2 \end{pmatrix}, \tag{7}$$

where σ_a^2 and σ_b^2 are the variances of the estimators obtained from domains a and b, respectively. $\sigma_{a;ab(A)}$ is the covariance between the estimator obtained from domain a and the estimator obtained from domain ab in frame A. $\sigma_{b;ab(B)}$ is the covariance between the estimator obtained from domain b and the estimator obtained from domain ab in frame B.

3.2.2 Fuller and Burmeister's estimator

In a dual frame survey, frame sizes N_A and N_B are such that:

- $N_A = N_a + N_{ab}$;
- $N_B = N_b + N_{ab}$.

Fuller and Burmeister's [11] proposed an estimator that uses sample information from the frames to estimate N_{ab}. Let

$$\hat{N}_{ab}^A = \sum_{k \in S_{ab}^A} \frac{1}{\pi_k^A} \qquad and \hat{N}_{ab}^B = \sum_{k \in S_{ab}^B} \frac{1}{\pi_k^B} \tag{8}$$

be estimators for $\hat{N}_{ab}$ based on frame A and B sample information respectively. Then, Fuller and Burmeister's estimator is given by

$$\hat{t}_{FB} = \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \sum_{k \in S_b} \frac{y_k}{\pi_k^B} + \sum_{k \in S_{ab}} y_k^* + p_2\left(\hat{N}_{ab}^A - \hat{N}_{ab}^B\right), \tag{9}$$

where

$$y_k^* = \begin{cases} p_1 \dfrac{y_k}{\pi_k^A}, & \text{if } k \in S_{ab}^A \\ (1-p_1)\dfrac{y_k}{\pi_k^B}, & \text{if } k \in S_{ab}^B \end{cases} \tag{10}$$

p_1 and p_2 are weighting constants ($0 \leq p_1, p_2 \leq 1$) chosen to minimize the variance of $\hat{t}_{FB}$.

In agricultural surveys, Fuller and Burmeister's estimator can be promptly applied accomodating complex sampling design in each frame, if necessary.

The approximated variance components of this estimator are

$$\mathbf{d} = \begin{bmatrix} 1 \\ p_1 \\ 1 \\ (1-p_1) \\ p_2 \\ p_2 \end{bmatrix} \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_a^2 & \sigma_{a;ab(A)} & 0 & 0 & 0 & 0 \\ \sigma_{a;ab(A)} & \sigma_{ab(A)}^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_b^2 & \sigma_{b;ab(B)} & 0 & 0 \\ 0 & 0 & \sigma_{b;ab(B)} & \sigma_{ab(B)}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{N_{ab(A)}}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{N_{ab(B)}}^2 \end{pmatrix}, \tag{11}$$

where σ_a^2 and σ_b^2 are the variances of the estimators obtained from domains a and b, respectively. $\sigma_{a;ab(A)}$ is the covariance between the estimator obtained from domain a and estimator obtained from domain ab in frame A. $\sigma_{b;ab(B)}$ is the covariance between the estimator obtained from domain b and the estimator obtained from domain ab in frame B. $\sigma_{N_{ab(A)}}^2$ and $\sigma_{N_{ab(B)}}^2$ are the variances of the estimators $\hat{N}_{ab}^A$ and $\hat{N}_{ab}^B$, respectively.

3.2.3 Bankier's estimator

Bankier [1] estimator relies on a set of sampling weights that allows its estimator to be written as the sum of only two Horvitz-Thompson estimators, each one covering the sample data from one of the frames. Bankier's estimator can also be written in the general form (1) as shown below:

$$\hat{t}_B = \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \sum_{k \in S_b} \frac{y_k}{\pi_k^B} + \sum_{k \in S_{ab}} \frac{y_k}{\pi_k^A + \pi_k^B} \tag{12}$$

Bankier's estimator can be extended to the use of multiple frames in a more simpler way than Hartley's. In its original paper, Bankier discussed the case were stratified probability samples are selected from frames A and B. Considering again frame A as an area frame, stratification of segments by land use is a common situation found in agricultural surveys. Stratification can also be applied to list frame B when auxiliary information is available. Applying Bankier's estimator in such cases have the appealing of producing estimates based on a simpler estimator in the sense that it combines estimates obtained from frames A and B separaetely.

The approximated variance components of Bankier's estimator are

$$\mathbf{d} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_B^2 \end{pmatrix}, \tag{13}$$

where σ_A^2 and σ_B^2 are the variances of the estimators obtained from domains a and b, respectively.

3.2.4 Skinner and Rao's estimator

Skinner and Rao (1996) proposed a pseudo maximum likelihood estimator (PML) that uses a single set of weights in each frame. This estimator can be expressed as:

$$\hat{t}_{SR} = \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \sum_{k \in S_b} \frac{y_k}{\pi_k^B} + \sum_{k \in S_{ab}} \gamma y_k^*, \tag{14}$$

where

$$y_k^* = \begin{cases} \dfrac{y_k}{\pi_k^A}, & \text{if } k \in S_{ab}^A \\ \\ \dfrac{y_k}{\pi_k^B}, & \text{if } k \in S_{ab}^B \end{cases} \qquad \gamma = \frac{\hat{N}_{ab,SR}}{\left(\pi_k^A \hat{N}_{ab}^A + \pi_k^B \hat{N}_{ab}^B\right)}. \tag{15}$$

$N_{ab,SR}$ is the smallest root of the quadratic equation $\alpha_1 x^2 + \alpha_2 x + \alpha_3 = 0$, where $\alpha_1 = n_A + n_B$, $\alpha_2 = n_A N_B + n_B N_A + n_A \hat{N}_{ab}^A + n_B \hat{N}_{ab}^B$ and $\alpha_3 = n_A \hat{N}_{ab}^A N_B + n_B \hat{N}_{ab}^B N_A$.

The approximated variance components of Skinner and Rao's estimator are

$$\mathbf{d} = \begin{bmatrix} 1 \\ 1 \\ \left(\dfrac{\pi_A}{\pi_A + \pi_B}\right) \\ \left(\dfrac{\pi_B}{\pi_A + \pi_B}\right) \\ \mu_a \\ \mu_b \\ \mu_{ab} \end{bmatrix} \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_a^2 & \sigma_{a;ab(A)} & 0 & 0 & 0 & 0 & 0 \\ \sigma_{a;ab(A)} & \sigma_{ab(A)}^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_b^2 & \sigma_{b;ab(B)} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{b;ab(B)} & \sigma_{ab(B)}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{N_{ab(A)}}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{N_{ab(B)}}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{N_{abPML}}^2 \end{pmatrix}, \tag{16}$$

μ_a, μ_b and μ_{ab} are the population means of the domains a, b and ab, respectively. The descriptions about the variances σ_a^2, $\sigma_{ab(A)}^2$, σ_b^2, $\sigma_{ab(B)}^2$, $\sigma_{N_{ab(A)}}^2$, $\sigma_{N_{ab(B)}}^2$ and the covariances $\sigma_{a;ab(A)}$ and $\sigma_{b;ab(B)}$ were presented previously.

Skinner and Rao's estimator can promptly accomodate complex sampling designs with high efficiency.

Given a set of two or more agricultural frames described in Table 1, and a dual frame sampling design that incorporate complex features such as stratification and multiple stages, there are situations where more than one estimator can be feasibly applied. In such cases, running small Monte Carlo simulation experiments based on available past surveys information can guide the choice. A good choice of estimator should take into consideration not only statistical performances but also simplicity. In this sense, screening type estimators are always good candidates and should not be discarded.

Although with great potential to generate agricultural statistics with high precision, the adoption of dual frame should be made upon studies to evaluate the cost-benefits in terms of gain in precision and accuracy, versus increase in complexity by dealing with two frames, when compared with a single frame survey option. Lohr and Rao [16] also call attention to the fact that, when using a multiple frame survey, care should be taken if samples selected from different frames employ different data collection instruments, such as different questionnaries. In such cases, there is potential to introduce bias at the estimates. Moreover, since the dual frame estimator's rely on the identifiability requirement, the estimates can be biased if problems with misclassification of sampled units in the correct domains happen.

4 Record linkage methods

In this paper, we could see that information raised by matching records from different frames can be helpful in many ways to improve statistical estimators, such as building good quality frames by combining information from several sources, helping frame maintainance over time, or identifying overlapping units in a dual frame estimation setup.

The goal of linking area and list frames at the frame component unit level leads to the subject of record linkage. A brief review of record linkage theory is introduced next. Several authors have been contributing to the development of a theory about the problem of record matching.

In order to have a basic idea about the type of problem treated by record linkage theory, we introduce a description based on the Fellegi and Sunter [10] paper. For simplicity, consider the availability of two list frames A and B. The first one has N_A records, and the second, N_B. In a simple scenario, suppose both frames are of type 1 (see Table 1), where holders addresses are the component units. Then, their records (items) are street names and numbers, for instance. We assume a general dual frame scenario, as in Figure 1, with domains a, b, and ab. Then the target population U is such that $U = a \cup b \cup ab$.

Let a record be labelled i or j if it takes the i-th or the $j-th$ value in A and B respectively. Given an enquiry record $i \in A$ and a particular file record $j \in B$, the problem of stablishing a linkage between them consists of evaluating the evidence that the pair (i,j) relates to the same element, in this example, the same address. We will represent such a matching by the expression $i = j$. In this case, the set of all records such that $i \in A$, $j \in B$, and a matching is stablished ($i = j$) coincides with the overlapping domain ab. The set of all records $i \in A$ such that $i \neq j$ is the domain a, and the set of all records $j \in B$ such that $i \neq j$ is the domain b.

In the search for a decision rule that stablish a link between records, one has to deal with the fact that frame records are subject to errors, and that sometimes only partial information is available, such as in cases where addresses numbers are missing.

The distribution of errors can be studied using simulation with artificial frames where one knows the right linkings between pairs. Let the distribution of pairs of records (i, j) given that they correspond to the set of addresses identified by both frames be

$$P\left(i,j|\text{match}\right) = p_{ij}, \qquad i,j = 1,2,\ldots,n \tag{17}$$

paint

Asume that the matrix of values p_{ij} is symmetric and denote the marginal distribution, by p_i. Then,

$$p_{ij} = p_{ji} \qquad\qquad p_i = \sum_j p_{ij} \tag{18}$$

The record linkage theory aims to find a linkage rule L, defined as a mapping from all the possible pairs (i,j) onto a set of random decision functions D that describes the result of a matching. For example, D can assume values d_1, d_2 or d_3, where d_1 represents the decision that there is a link between i and j, i.e. $i = j$; d_2 represents an inconclusive decision, and d_3 says there is no match, i.e. $i \neq j$. The chosen decision can imply in two types of error:

- Type I Error: $D(i,j) = d1 \mid d_3$ is true;
- Type II Error: $D(i,j) = d3 \mid d_1$ is true.

Fellegi and Sunter established conditions under which a linkage rule L_0 is optimal in the sense that if L^* is any competitor of L_0 having the same Type I and Type II error probabilities, than conditional probabilities (either on the domain sets a, b, or on the overlapping domain ab) of not making a decision under rule L^* are always greater than under L_0.

Let M denote the set of all pairs (i,j) such that a matching holds ($i = j$), and M^c denote its complement, i.e. the set of all pairs (i,j) such that ($i \neq j$). Then, the Fellegi and Sunter optimal rule L_0 is written as a function of

$$R = \frac{P(D(i,j) \mid (i,j) \in M)}{P(D(i,j) \mid (i,j) \in M^c)} = \frac{p_{ij}}{p_i p_j}. \tag{19}$$

If R assumes a value greater than a certain upper limit u, L_0 indicates $D = d_1$; if R assumes a value lower than a certain lower limit l, L_0 indicates $D = d3$; Values of R within the interval (l,u) lead to $D = d_2$.

Winkler and Thibaudeau [21] described an application of the Fellegi-Sunter model to the 1990 US decenial population census. Yitzkov and Azaria [22] application of record linkage theory to the Israels Central Bureau of Statistics project on making a transition from a traditional census to an Integrated Census where the source for population counts is obtained from administrative files. In further versions of this paper we intend to describe studies of the potential use of record linkage methods ingricultural surveys.

5 Conclusion remarks and Recommendations

In this paper, a preliminary review on the literature concerning methods for linking area and list frames was presented. Although this is an on going process, some conclusion remarks can be outlined.

Linking information from different registers is an essential task for building list frames with full coverage. The incresing computational ability to handle massive data sets make real

the possibility of exploring administrative data as one of these source registers. However, this approach should be taken only if the different sources contribute with essential information to the frame and the record matching gives extremely reliable results. The cost of such building process should also be evaluated, not only in terms of complexity, but also with respect to the potential impact of matching errors on statistical estimates. Studies aiming to assess such cost are reccommended.

An alternative to dealing with a list frame building process is to use available incomplete list frames in a multiple frame approach. If their simultaneous use is not sufficient to guarantee full population coverage, an area frame should be added to the multiple frame survey in order to avoid bias. Since using multiple frames represent a significant increase in complexity, the design of a dual frame survey using an area and a list frame can be a good compromise. In this case, using an area frame has the further advantage of allowing objective assessment of land characteristics, such as cultivated area, if needed.

The choice of the dual frame sample design will depend on the types of frames available (see Table 1). Frames with clustered elements may need a multiple stage design before identifiability requirement hold. In addition, more than one dual frame estimator can be a candidate to be applied. In such cases we reccomend running Monte Carlo simulation experiments based on available agricultural information to guide the choice.

A final remark concerning the use of record linkage methods emphasizes its potential to improve the building process of good quality list frames, as well as to minimize identifiability problems in dual frame designs. Therefore, studies focusing on the application of record linkage methods to agricultural surveys should be encouraged.

References

[1] Bankier, M. D. (1986) *Estimators based on several stratified samples with applications to multiple frame surveys.* Journal of the American Statistical Association, 1986..

[2] Benedetti, R., Bee, M., Espa, G. and Piersimoni, F. (2010). *Sampling Methods for Agricultural Surveys.* Wiley, Chichester, UK.

[3] Carfagna, E. (2001) Multiple frame sample surveys: advantages, disadvantages and requirements. In International Statistical Institute, Proceedings, Invited papers, International Association of Survey Statisticians (IASS) Topics, Seoul, August 22-29, 2001, 253270.

[4] Carfagna, E. (2007) A comparison of area frame sample designs for agricultural statistics. Bulletin of the International Statistical Institute, 56th Session (Lisbon), Proceedings, Special Topics Contributed Paper Meetings (STCPM11), Agricultural and Rural Statistics, 1470, 2229 August 22-29, pp. 14.

[5] Carfagna E. and Carfagna A. (2010). *Alternative sampling frames and administrative data; which is the best data source for agricultural statistics?* . in Benedetti, Bee, Espa, Piersimoni (Editors), Agricultural Survey Methods, Wiley, Chichester, UK.

[6] DOrazio, M., Di Zio, M. and Scanu, M. (2006) *Statistical Matching: Theory and Practice.* Chichester: John Wiley & Sons, Ltd.

[7] FAO. (1989). *Sampling Methods for Agricultural Surveys.* FAO Statistical Development Series No. 3. Rome, FAO.

[8] FAO. (1996). *Multiple Frame Agricultural Surveys, Volume 1.* FAO Statistical Development Series No. 7. Rome, FAO.

[9] FAO. (1998). *Multiple Frame Agricultural Surveys, Volume 2.* FAO Statistical Development Series No. 10. Rome, FAO.

[10] Fellegi, I. P., and Sunter, A. B. (1969), A Theory for Record Linkage, *Journal of the American Statistical Association, 40, 1183-1210.*

[11] Fuller, W. A. and Burmeister, L. F. (1972). *Estimators for samples selected from two overlapping frames.* Proceeding of the Social Statistics Association - ASA.

[12] Haines, D. E. and Pollock, K. H. (1998). Estimating the number of active and successful bald eagle nests: an application of the dual frame method. *Environmental and Ecological Statistics.* **5**, 245–256.

[13] Hartley, H. O. (1962). Multiple frame surveys.. *Proceedings of the Social Statistics Section, American Statistical Association* pp.203–206.

[14] Horvitz,D. G.; Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Associtation, 47, 663, 685.*

[15] Lavallee, P. (2005). Quality indicators when combining survey data and administrative data. In Proceedings of Statistics Canada Symposium 2005: Methodological Challenges for Future Information Needs. http://www.statcan.gc.ca/pub/11-522-x/11-522-x2005001- eng.htm..

[16] Lohr, S. and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Associtation, Vol.95,No.449. pp.271-280.*

[17] Shin, H; Molinari, N. and Wolter, K. (2008). *A dual-frame design for the National Immunization Survey.* JSM Proceedings, Alexandria, VA: American Statistical Association..

[18] Selander, R., Svensson, J., Wallgren, A. and Wallgren, B. (1998). *How should we use IACS data? Statistics Sweden..*

[19] Skinner, C. and Rao, J. N. K. (1996). *Estimation in dual frame surveys with complex designs.* Journal of the American Statistical Association.

[20] Wallgren, A. and Wallgren, B. (1999). *How can we use multiple administrative sources? Statistics Sweden.* .

[21] William E. Winkler and Yves Thibaudeau (1987). An Application Of The Fellegi-Sunter Model Of Record Linkage To The 1990 U.S. Decennial Census. U.S. Decennial Census Technical report, US Bureau of the Census.

[22] Yitzkov , T. and Azaria, H. Record Linkage in an Integrated Census