Technical Report Series GO-08-2015

Technical Report on

Linking Area and List Frames in Agricultural Surveys

October 2015

Technical Report on

Linking Area and List Frames in Agricultural Surveys

Table of Contents

Figures

Tables

Acronyms and Abbreviations

CSA	Central Statistical Agency (Ethiopia)
DFID	Department for International Development
EA	Enumeration Area
FAO	Food and Agricultural Organization of the United Nations
GMHT	Generalized Multiplicity - Horvitz - Thompson
GMRE	Generalized Multiplicity Ratio Estimators
IBGE	Brazilian Institute of Geography and Statistics
ICAS	Integrated Administrative and Control System
MPPS	Multivariate Probability – Proportional – to Size
NSS	National Statistical Systems
PPS	Probability – Proportional – to Size
SAC	Scientific Advisory Committee
SRS	Simple Random Sampling
STCPM	Special Topics Contributed Paper Meetings

Preface

This Technical Paper on **Linking Area and List Frames in Agricultural Surveys** was prepared within the framework of the Global Strategy to Improve Agricultural and Rural Statistics. The Global Strategy is an initiative endorsed in 2010 by the United Nations Statistical Commission, to provide a framework and a blueprint to meet current and emerging data requirements and the needs of policymakers and other data users. The Global Strategy's goal is to contribute to greater food security, reduced food price volatility, higher incomes and greater well-being for rural populations through evidence-based policies. The Global Strategy's Global Action Plan is centred upon 3 pillars: (1) establishing a minimum set of core data (2) integrating agriculture into National Statistical Systems (NSSs) and (3) fostering the statistical system's sustainability through governance and statistical capacity building.

The Action Plan to Implement the Global Strategy includes an important Research Programme, which addresses methodological issues in improving the quality of agricultural and rural statistics. It is envisaged that the Research Programme will devise scientifically sound and cost-effective methods that can be used as reference material when preparing practical guidelines for country statisticians, training institutions, consultants, etc.

To enable countries and partners to benefit from the results of the Research activities at an early stage, a **Technical Reports Series** has been established, to widely disseminate existing technical reports and advanced draft guidelines and handbooks. This will also provide opportunities to receive early and further feedback on the papers.

The Technical Reports and the draft guidelines and handbooks published in this Technical Report Series have been prepared by Senior Consultants and Experts and reviewed by the Scientific Advisory Committee (SAC)[1] of the Global Strategy, the Research Coordinator of the Global Office and other independent Senior Experts. For certain research topics, field tests will be organized before the final results are published in the relevant guidelines and handbooks.

[1] The SAC is composed of 10 renowned senior experts in various fields relevant to the Research Programme of the Global Strategy and are selected for a term of two years. The SAC members who reviewed this report are: Vijay Bhatia, Seghir Bouzaffour, Ray Chambers, Jacques Delincé, Cristiano Ferraz, Miguel Galmes, Ben Kiregyera, Sarah Nusser, Fred Vogel, and Anders Walgreen.

In a general framework, agricultural surveys use area frames and list frames as the main sampling frames to identify and access the elements of the target population. On the basis of the sampling frame and the type of auxiliary information available, a probability sample is statistically designed and selected, data is collected and analysed, and estimates are produced. The accuracy and precision of official statistics may be affected by any step of the survey process, but frame quality has a major effect on the efficiency of statistical planning and analysis as a whole. Frame-building processes depend on the resources available. While area frames may take advantage of the most recent technological advances in satellite imagery and global positioning systems, list frames can be built on the basis of information from administrative records and inherited data from recently implemented agricultural or demographic censuses.

This Technical Report on *Linking Area and List Frames in Agricultural Surveys* is the result of a comprehensive literature review on the subject and further methodological developments. The Report introduces and discusses the problem of improving the quality of agricultural statistics by exploring methods to maximize the use of available frames, focusing on how to combine information from area frames and list frames in agricultural surveys. The Report reviews basic concepts relating to area and list frames, highlighting their main advantages and disadvantages when used in a single-frame survey approach. The possibility of linking the two types of frames at the design and estimation stages of a survey is considered, and the gains involved in each strategy discussed. It is argued that linking area and list frame information through a multiple frame approach is the best option for improving the quality of agricultural statistics.

This Report will supplement the Handbook on Master Sampling Frame for Agriculture, which is currently under preparation.

Acknowledgments

The Technical Paper on *Linking Area and List Frames in Agricultural Surveys* was prepared by Cristiano Ferraz, Professor at the Federal University of Pernambuco, Brazil, Department of Statistics, with the guidance and supervision of Naman Keita, Research Coordinator, and Michael Rahija, Research Officer, of the Global Office of the Global Strategy to Improve Agricultural and Rural Statistics (FAO).

The author wishes to thank Hemilio Coelho (Federal University of Paraíba), Carla Monteiro (Federal University of Pernambuco), Elisabetta Carfagna (University of Bologna) and Fulvia Mecatti (University of Milano-Bicocca) for their kind and invaluable contributions.

Valuable input and comments were provided at different stages by the SAC members and by participants in the Expert Meeting on Master Sampling Frame held in Rome on 30-31 October 2014[2].

This publication was prepared with the support of the Trust Fund of the Global Strategy, funded by the UK's Department for International Development (DFID) and the Bill & Melinda Gates Foundation.

[2] Luis Ambrosio, University of Madrid; Robert Arcaraz, Ministry of Agriculture – France; Marino Barrientos, University of San Carlos – Guatemala; Ambika Bashyal, Central Bureau of Statistics – Nepal; Roberto Benedetti, Università degli Studi G. d'Annunzio Chieti e Pescara – Italy; Flavio Bolliger, Instituto Brasileiro de Geografia e Estatistica – Brazil; Loredana di Consiglio, ISTAT – Italy; Javier Gallego, EU/JRC; Luis Iglesias, University of Madrid – Spain; Dalisay (Dax) Maligalig, Asian Development Bank; Sebastien Manzi, Institut National des Statistiques Rwanda; Giovanna Ranalli, University of Perugia – Italy; Michael Steiner, USDA/NASS; Aberash Tariku, Central Statistical Bureau of Ethiopia; Xinhua Yu, National Bureau of Statistics China; Nancy Chin, FAO; John Latham, FAO; Naman Keita, FAO; Eloi Ouedraogo, FAO; Mukesh Srivastava, FAO.

1

Introduction

Quality agricultural and rural statistical data play a significant role in helping politicians and decision makers diagnose scenarios and create and evaluate public policies. The current global challenges of increasing food production, reducing poverty and maintaining environmental sustainability reinforce the need for and importance of building efficient national statistical systems, capable of providing official statistics on agriculture that are timely, accurate and comparable across countries.

The experiences of different countries show that agricultural censuses, sampling surveys, and administrative registers are the main sources used when generating official data related to agriculture. The feasibility of implementing sampling surveys between censuses on a periodical basis endows these surveys with a major role in the statistics production system. An essential tool in conducting these surveys is the sampling frame. Frames are used to identify and access the elements of the survey's target populations. As for the nature of their components, the frames commonly adopted in agricultural surveys are either list frames or area frames. The decision to adopt one or the other type depends on the building and maintenance costs of each, balanced with their capacity to provide estimates with acceptable levels of precision and accuracy.

Since Hartley's groundbreaking paper (Hartley 1962), the possibility of using both types of frame has received attention in terms of methodology, including in FAO's two-volume publication on Multiple Frame Agricultural Surveys (FAO 1996, 1998). Indeed, combining the information from the two types of frames can be strategic in providing a survey design scenario of higher quality: there would be greater chances of auxiliary information being available, the impact of outliers upon estimates' precision would be reduced and there would be fewer chances of nonsampling errors arising in multipurpose surveys – a notable characteristic of surveys for agriculture and rural assessment.

This Technical Report introduces and discusses methods for linking area frames and list frames, and considers the effects of these linkages during a survey's design and estimation stages. A literature review of the subjects involved in the process of linking information from both frames is provided, together with a

description of the statistical methods applied. General descriptions of the topics are then set out, followed by specific comments on applying the methods to agricultural surveys and the feasibility of their adoption by developing countries. In addition to this Introduction, this Report contains seven sections. Section 2 introduces the fundamental concepts relating to frames used in agricultural surveys and the role of administrative data in generating and improving them, within the context of a single frame survey. Section 3 describes the problem of drawing inferences based on a multiple-frame source, emphasizing the possibility of linking frames at the design or estimation stages of a sample survey. In Section 4, the usage of record linkage methods to link area frames with list frames at the design stage is explored. Section 5 describes the multiple frame sample design and estimation theory, focusing on the dual frame scenario that can accommodate linking information from area and list frames during the estimation stage of an agricultural survey. Sections 6 and 7 explore the design aspects of dual frame surveys, examining sample size allocation and the use of auxiliary information respectively. Finally, in Section 8, conclusions and recommendations are given.

2

Basic frame concepts

A *sampling frame* (or simply a *frame*) is a fundamental concept in survey sampling, as it plays a major role in the quality of the statistical inference. A frame may be defined as a reference system composed of a set of materials, devices or coordinates that enables selection of a sample. This reference system must be capable of providing the information necessary to access the elements of the population of interest (Särndal et al. 1992). In agricultural surveys, typical examples of populations of interest – also called target populations – are the set of all the holdings (farms) in a country and the set of all segments of area covering a country.

Probability sampling schemes are applied to select the frame's component units (sampling units) that lead to the identification of elements or sets of elements (clusters) within the target population. In the latter case, if necessary, additional material must be used to identify the elements in each cluster and to select a subsample thereof, following a further sampling scheme. The subsampling process may continue as required. The choice of sampling scheme depends on the nature of the frame component units in use, as well as on the nature of any auxiliary information that may be available. Therefore, the capability of the chosen sampling design (sampling scheme and estimator) to provide accurate and precise estimates for the parameters of interest depend on the quality of the information provided by the frame itself (or by the set of frames, in the case of subsampling design). Ideally, the frame should provide full coverage and uniquely identify all target population elements. Achieving appropriate coverage and identification, however, often requires information to be updated frequently. Auxiliary information may also be available for frames; failing to taking advantage of this leads to the use of less statistically efficient estimators. On the other hand, failing to properly identifying restrictions to frame coverage level and unique identification can lead to bias and variance inflation of estimates.

Traditionally, in cases where the frame component units are segments of land, the frame is called an *area frame*, and the sampling procedure adopted is generally identified as area sampling. In these cases, when information on the area measure of each segment is available, a commonly chosen probability

sampling scheme applied to area frames is sampling with probability proportional to the area measure. However, area frame component units are not restricted to segments of areas. Sampling from an area frame can also be accomplished by randomly selecting points in a coordinate system and then observing a segment of area around each sampled point (Gallego 2013). In several circumstances, on the other hand, the frame components are simply lists of addresses for holders or farmers, justifying the so-called general expression of *list frames*.

In time, the existing types of area frames and list frames have been improved. Advances in technology and computing resources have enabled the construction of more efficient area frames, supported by high quality satellite images, as well as more informative list frames, with data combined from different registry sources. However, area frames and list frames may be described briefly as in subsections 2.1 and 2.2 below. The role of administrative data in building and maintaining quality list frames is also considered below (Section 2.3).

2.1. List Frames

Usually consisting of a list of holdings (farms) or holders' addresses, list frames are the most common type of frame in agricultural probability sample surveys. They may be built on the basis of information collected from the most recent agricultural or population census, administrative data, previous surveys or a combination of several data sources. If their component units are clusters, multistage sampling schemes with further frames to refine the identification of the clusters' elements may be necessary to reach the target population.

Auxiliary information may be available in list frames, which would enable the use of efficient sampling schemes such as stratified sampling, probability-proportional-to-size sampling or even both of these, as well as that of calibration and regression-type estimators.

Although sampling costs depend on the choice of the sample design, sampling and identification of reporting units in agricultural surveys have a relative low cost when using list frames, as the sampled farmers names and addresses are listed, either as the one stage or multistage final sampling units, becoming promptly accessible for the field work. The sampling and identification of reporting units in agricultural surveys have a relatively low cost when using list frames, because the sampled farmers' names and addresses are listed either as the one-stage or multistage final sampling units, and are thus promptly accessible for the fieldwork.

However, list frames are subject to rapid degeneration over time, which may lead to problems of undercoverage and obsolete information if they are not properly maintained. It should also be sought to avoid problems with duplicate records.

Countries with advanced information systems can take advantage of information from several sources, including administrative data, to build efficient list frames. The challenge of achieving high rates of successful matching, however, entails one of the main costs of the building process. Further studies to estimate the magnitude of these costs are necessary. D'Orazio et al. (2006) review a series of methods that aim to match records between files. The potential impact of matching process errors to statistical estimates is also a subject upon which further studies would be required.

2.2. Area Frames

Area frames may consist of an infinite set of points or of a finite set of area segments. The segments of an area composing an area frame can be determined in different ways: they may be established by reference to identifiable physical boundaries such as rivers, roads etc., by means of a squared grid of map coordinates or by making their limits coincide with those of agricultural holding lands (FAO 1996). When the segment does not coincide with the boundaries of a holding, a tract must be defined. Segments are then subdivided into non-overlapping tracts, in which a tract is the part of a holding that is found within the limits of a segment, or a piece of land that does not belong to any holding. A holding comprises at least one tract. Tracts are observational units. Gallego et al. (1994) and Gallego (1995, 2013) provide information on sampling points from an area frame for agricultural surveys.

Area frames have the advantage of providing full coverage of the target population, are duplication-free and remain up-to-date for a long time. In addition, they are ideally suited to the generation of estimates of parameters relating to land areas, such as a total cultivated area, as they enable objective measures to be taken on the ground. On the other hand, although their costs are falling, they remain expensive to use for drawing samples. In addition, Carfagna (2004) notes that the presence of outliers in samples from area frames has a considerable impact on estimates.

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

Table 1 – Types of area and list frame suitable for agricultural surveys

Frame type	Frame description	Unit component	Unit type
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
6	Area frame	Point	Area around the point

2.3. Frames and Administrative Data

Adequate list frames can be built on the basis of information collected in succesfully executed agricultural and population censuses. However, maintaining their adequacy over time depends on the availability of an efficient information system that can update the data continuously. Due to the constant advances in information technology, it is feasible to explore and evaluate the potential role that administrative registries could perform in feeding these systems, as well as the possibility of using surveys that are based only on administrative data to generate agricultural statistics. Selander et al. (1998) and Wallgren and Wallgren (1999), investigated the possibility of producing statistics on crops and livestock using the Integrated Administrative and Control System (ICAS), a European system of information relating to agricultural subsidies. Carfagna and Carfagna (2010) analysed the advantages and disadvantages of producing agricultural statistics based on administrative data instead of survey samples. Designing surveys based only on administrative data differs in several respects from designing surveys based on sample data. In sample surveys, the data collection process takes place after certain aspects of sampling design, such as target population definition and parameters of interest, have been defined. In administrative data surveys, the data collection process precedes these stages, because the data have already been collected. Although the use of administrative data to produce agricultural statistics entails a series of challenges concerning proper coverage, managing data quality and identifying adequate variables of interest, their joint use with survey data holds great potential for the improvement of the design of agricultural surveys, as demonstrated by Lavallée (2005). Identifying methods to improve the matching of several sources of data at the level of frame component units may support the

feasibility of good quality frames, not only in terms of coverage rates but also of providing access to auxiliary information.

When dealing with area frames, stratification is often employed, and refreshing the information necessary for this task requires efforts to raise up-to-date and relevant auxiliary data. The possibility of using administrative records to raise such data should be evaluated carefully. Carfagna (2007) describes the results of the 2001 AGRIT project, which assessed efforts to change from an area frame with clusters to an area frame with uncluttered points, using ICAS data to refresh stratification variables. The 1.2 million points of the new sampling frame were stratified according to the following classes: arable land, permanent crops, permanent grass, forests, isolated trees and rural buildings, and other (artificial areas, water, etc.). The study concluded that it would not be worthwhile to update the area frame feature annually, as would be required by such a shift, because the accuracy of the land classification assessed by means of a sample was low.

This brief description of list frames, area frames and administrative data draws attention to the challenges that must be faced in using each as a sampling frame or even as an option to a survey (as the case of administrative data). If, on one hand, using administrative data instead of a survey may pose serious limitations, on the other these data can provide valuable information for the construction of efficient list frames. Quality list frames may be available on the basis of recent population censuses, but only for a short period of time. The issue of maintaining the frame information up-to-date will soon resurface; administrative records may be of assistance in this respect.

However, efforts to build and maintain a list frame based on different sources of data may not be sufficient to guarantee coverage of the full population. In these cases, although an area frame can be adopted to avoid bias, unique disadvantages (as mentioned above) still arise. Adopting a dual frame approach – in which the available (incomplete) list frame and the area frame are used simultaneously but independently of one another – may be a good compromise, to avoid incurring the disadvantages deriving from using list and area frames alone while still being able to exploit the best characteristics of each. In addition, a dual-frame survey provides flexibility in the choice of statistical sampling designs for each frame. The dual-frame idea can be generalized to a multiple frame survey, in which more than two frames simultaneously cover the same target population to support a sample survey design.

3

Multiple frames for agricultural surveys

Brazil and Ethiopia are some of the countries engaged in the composition of several frames into a master frame for integrating agricultural and population surveys. During the Sixth International Conference on Agricultural Statistics, held in Rio de Janeiro, Brazil, in 2013, statisticians from both countries presented a brief talk on the subject.

The Central Statistical Agency of Ethiopia (CSA) is responsible for conducting agricultural surveys in the country. It collects data on subjects such as cultivated area and production by crop type, land use and agricultural practices (Abaye 2013). In the survey design, the Enumeration Areas (EAs), defined for the population and housing census, are used as primary sampling units. Then, a listing of households found in each sampled EA is carried out, thus producing a frame list of households to be selected as secondary sampling units.

To improve the quality of the agricultural data, the CSA is currently studying the possibility of adopting the following composition of area and list frames: an area frame sampling is conducted with EAs as primary sampling units, and segments having a size of 40 hectares as secondary sampling units, stratified by land cover classification. In addition, data from commercial farms are also collected, using a list frame. To improve estimates, information from area and list frames should be combined.

The Brazilian Institute of Geography and Statistics (IBGE) is currently studying survey design options to implement the Brazilian Agricultural Survey System. The sampling frame for this system is a composition of area and list frames aiming to provide coverage of the target population of the country's agricultural establishments. As per Santos et al. (2013), "[*e*]*stablishments where the production is higher when compared to others are likely to be selected in the list frame while small establishments are going to be selected using the area frame. In the list frame there is a stratification by economic activity and the magnitude of the establishment. In some cases there are units selected with probability one. Others are selected using simple random sampling without replacement*."

The examples from both countries illustrate efforts to use several data sources in a master frame for agricultural surveys. Linking information from different sources may be performed at either the design or the estimation stage of a survey. The strategy of attempting to link data from several frames during a survey's design stage seeks to concentrate the efforts to build a single frame from multiple sources and conduct a single frame survey. One of the problems raised in this scenario is the appropriate matching of records among different registers, which leads to the subject of record linkage. However, in addition to the challenges ensuing when no perfect matching can be detected, the same problems of frame maintenance and costs apply to the survey resulting from this approach.

On the other hand, the strategy of linking information from different frames at the estimation stage derives from the idea of combining information from each frame without any need to match records before the survey data collection process. This strategy provides flexibility, as it enables a survey sampling to be designed from each frame independently, and is less prone to error than the strategy of linking data at the design stage. The multiplicity estimator proposed by Mecatti (2007) offers the theoretical foundation for inference with multiple frames.

Following the notation proposed by Mecatti and Singh (2014), let $U_1, U_2, \ldots, U_Q$ denote the collection of frames the union of which is assumed to cover the target population $\mathsf{U} = \bigcup_{q=1}^{Q} U_q$. The frames can overlap with one another and some may even provide full population coverage. According to this approach, independent samples denoted by $S_1, S_2, \ldots, S_Q$ are selected from each of the Q frames without having to maintain the same probability sample design. Consider the goal of estimating the population total of a study variable:

$$t = \sum_{k \in \mathsf{U}} y_k \quad , \tag{1}$$

where y_k is the value of element k in the population, based on data coming from samples Q. The total t may be expressed as the sum of all overlapping frames, i.e.

$$t = \sum_{k \in \mathrm{U}} y_k = \sum_{q=1}^{Q} \sum_{k \in U_q} y_k \alpha_{q(k)}, \tag{2}$$

where $0 \le \phi_{q(k)} \le 1$, in general but not necessarily, and $\sum_{q=1} \phi_{q(k)} = 1$ indicates the multiplicity adjustment factors corresponding to the qth frame and the kth unit. Let $\delta_{k(q)}$ be a random variable that represents the sample membership indicator of unit k in the sample S_q from frame U_q. Thus, the generalized multiplicity-adjusted Horvitz-Thompson (GMHT) estimator is given by

$$\hat{t} = \sum_{q=1}^{Q} \sum_{k \in U_q} y_k \phi_{q(k)} \frac{\delta_{k(q)}}{E\left(\delta_{k(q)}\right)}. \tag{3}$$

The Horvitz-Thompson estimator is a particular exemplification of this estimator, when $Q = 1$ and $\phi_{q(k)} = 1$.

4

Linking frames at the survey design stage

Combining information that matches records from several sources at the survey design stage is one way to build and update a sampling frame. However, in the context of agricultural applications, this method should be used with caution, as it may be inefficient compared to a dual frame approach (described in Section 5 below) or even unfeasible.

Linking area and list frames at the frame component unit level is a challenge, as it depends on the existence of a linking connection between area units and list units before the survey fieldwork is performed. In some countries, population and agricultural censuses have generated a database with point locations that could provide such a link.

Whenever feasible, the problem of matching records from two frames leads to the subject of record linkage. A brief overview of record linkage theory is introduced below by way of illustration. Several authors have contributed to the development of a theory on the problem of record matching.

To obtain a basic idea of the type of problem treated by the theory, a description based on Fellegi and Sunter's paper (1969) is introduced. For the sake of simplicity, consider the availability of two list frames A and B. The first has N_A records and the second N_B records. Whenever necessary, the superscripts A or B indicate the frame from which the information is related. In a simple scenario, suppose that both frames are of Type 1 (see Table 1 above), in which the component units are addresses of holders. Thus, their records (items) are e.g. street names and numbers. It is assumed that the frames contain common elements, such as in Figure 1, with domains a, b, and ab. Thus, the target population U is such that $\mathsf{U} = a \cup b \cup ab$.

Let a record be labeled i or j if it takes the i-th or the j-th value in A and B respectively. Given an enquiry record $i \in \mathsf{A}$ and a particular file record $j \in \mathsf{B}$, the problem of establishing a linkage between these consists in evaluating the

evidence that the pair (i, j) is related to the same element – in this example, that they share the same address. This matching can be represented by the expression $i = j$. In this case, the set of all records such that $i \in \mathsf{A}$, $j \in \mathsf{B}$, and a matching is established ($i = j$) coincides with the overlapping domain ab. The set of all records $i \in \mathsf{A}$ such that $i \neq j$ is the domain a, and the set of all records $j \in \mathsf{B}$ such that $i \neq j$ is the domain b.

When seeking a decision rule that establishes a link between records, it is necessary to address the fact that frame records are subject to error, and that sometimes only partial information is available, such as when address numbers are missing.

The distribution of errors can be studied using simulation with artificial frames, if the correct linking between pairs is known. Let the distribution of pairs of records (i, j), given that they correspond to the set of addresses identified by both frames, be

$$P(i, j \mid match) = p_{ij}, \qquad i, j = 1,2,\ldots,n. \tag{4}$$

Assume that the matrix of values p_{ij} is symmetrical and denotes the marginal distribution with p_i. Then,

$$p_{ij} = p_{ji} \qquad\qquad p_i = \sum_j p_{ij} \tag{5}$$

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 may assume the values d_1, d_2 or d_3, in which 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 indicates there is no match, i.e. $i \neq j$. The chosen decision may imply 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, then the conditional probabilities (either in relation to the domain sets a, b, or to 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 optimal Fellegi and Sunter rule L_0 is written as a function of

$$R = \frac{P(D(i, j) \mid (i, j) \in M)}{P(D(i, j) \mid (i, j) \in M^c)} = \frac{p_{ij}}{p_i p_j}. \quad (6)$$

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 (1987) described an application of the Fellegi-Sunter model to the US's 1990 decennial population census. Yitzkov and Azaria (2003) apply record linkage theory to Israel's Central Bureau of Statistics' project on shifting from a traditional census to an integrated census, in which the source for population counts is obtained from administrative files.

When feasible, the process of matching information from area and list frames via record linkage methods may not be efficient, as the costs entailed by the task and the problems that may arise from non-conclusive results can all be avoided by using a dual-frame design for the survey.

5

Linking frames at the survey estimation stage

The literature features several dual frame estimators. Their differences lie essentially in how they make use of the information yielded by both frames. In this sense, dual-frame inference may be viewed as a form of linking two frames during a survey's estimation stage.

Let the target population U be covered by two frames A and B. As originally proposed by Hartley (1962), inference in dual frame surveys considers the broad scenario illustrated by Figure 1 below, in which three domains can be identified: $a = \mathsf{A} \cap \mathsf{B}^c$, $b = \mathsf{B} \cap \mathsf{A}^c$ and $ab = \mathsf{A} \cap \mathsf{B}$. There are two important requirements:

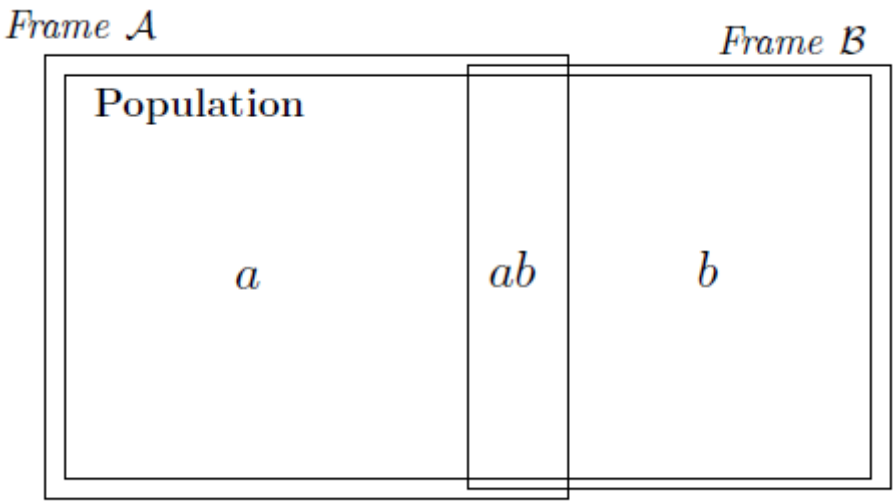
- completeness, and
- identifiability.

The completeness requirement is met when both of the frames used in the design provide full coverage for the target population, such that $U = \mathsf{A} \cup \mathsf{B}$. In this way, every element is listed in at least one of the frames. The identifiability requirement is fulfilled when, for any sampled element, it is possible to understand whether or not it belongs to one of the frames. Although the dual frame method can be generalized to multiple frames, dealing with more than two frames makes it more difficult to meet the necessary requirements. The multiplicity estimator (Mecatti and Singh 2014) offers a more flexible option, because it does not require identifiability as explained here, but only awareness of a multiplicity factor.

Dual-frame surveys are flexible enough to accommodate different types of frames. Figure 1 below could represent the use of two incomplete list frames. On the other hand, if Frame A is an area frame, and Frame B is a list frame, for example, then a particular case in which $\mathsf{B} \subset \mathsf{A}$ occurs, as illustrated by Figure 2 further below. In this case, using a dual-frame approach may lead to pecuniary savings if the costs of sampling from Frame B are lower than those

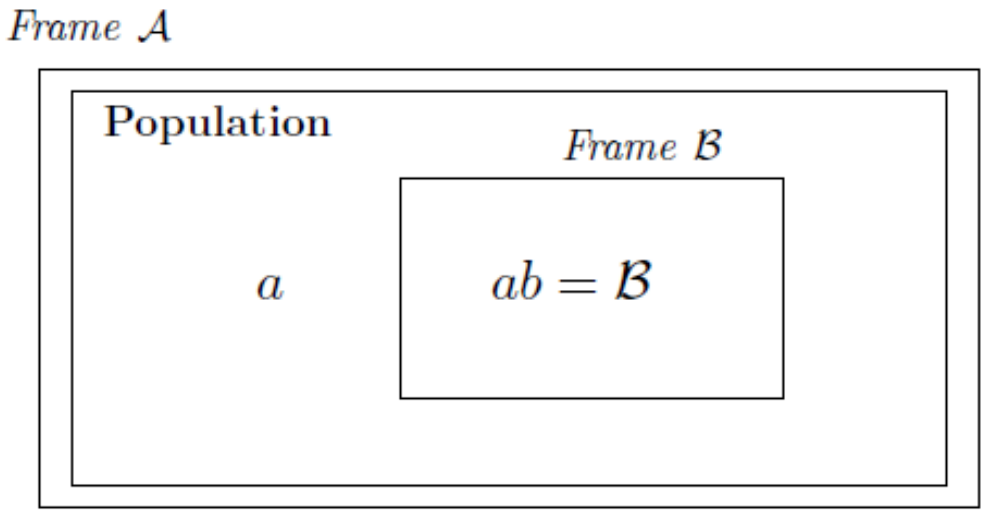
relating to Frame A; the dual frame approach may also improve accuracy, if auxiliary information is available from Frame B.

Figure 1 – A general dual frame scenario



The choice of an appropriate sampling design for each frame depends on the amount of information available and whether it is possible to identify the elements in the domains. If efficient linkage methods can be applied to the frames to match their records and identify the elements in each domain, the stratification of the population is feasible, and the 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 the construction of a unique frame, which unifies information from Frames A and B.

Figure 2 - A special dual frame scenario based on an area frame and a list frame



If 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:

$$\hat{t} = \hat{t}_a + \hat{t}_b + \hat{t}_{ab}, \tag{7}$$

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 domain totals. The variances depend on how each estimator deals with information raised from both frames. In general,

$$Var(\hat{t}) = Var(\hat{t}_a) + Var(\hat{t}_b) + Var(\hat{t}_{ab}) + 2[Cov(\hat{t}_a, \hat{t}_{ab}) + Cov(\hat{t}_b, \hat{t}_{ab})]. \tag{8}$$

Carfagna (2001) reviews the main advantages, disadvantages and requirements of dual frame designs. This Technical Report focuses on the four major estimators proposed in the literature and outlines their potential use with respect to the types of agricultural frame described in Table 1 above.

5.1. Dual Frame Estimators

Let π_k^A and π_k^B be the first-order inclusion probabilities for the 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 U_{ab} denote the set of population elements belonging to domain ab, while S_{ab}^A denotes the sample set of elements from U_{ab} selected from frame A. For instance, $\bar{y}_{ab}^B$ denotes the ab domain sample mean from frame B. Below, some dual frame estimators suggested in the literature will be introduced.

5.1.1. Hartley's estimator

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

$$\hat{t}_H = \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \sum_{k \in S_b} \frac{y_k}{\pi_k^B} + \sum_{k \in S_{ab}} y_k^*. \tag{9}$$

In Expression 9,

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

and p is a weighting constant ($0 \le p \le 1$) chosen to minimize the variance of the estimator $\hat{t}_H$. To apply the estimator, identifiability is crucial.

If necessary, although not as user-friendly as that shown above, Hartley's estimator can also be written using the multiplicity estimator notation to facilitate a link with multiple frame designs:

$$\hat{t}_H = \sum_f \sum_k d_{fk} y_k \phi_k^f = \sum_f \hat{Y}_f^*, \tag{11}$$

where $\hat{Y}_f^* = \sum_k d_{fk} y_{fk}^*$ and $y_{fk}^* = y_k \phi_k^f$. In this expression, ϕ is a weighting constant ($0 \le \phi \le 1$) chosen to minimize the variance of the estimator $\hat{t}_H$, such that $\phi_k^A = \psi$ and $\phi_k^B = 1 - \psi$ if the element k is in both frames, $k \in S_{ab}^A$ and $k \in S_{ab}^B$ respectively, $0 \le \psi \le 1$, and $\phi_k^f = 1$ if element k is in only one frame. The weights for each frame $f = A, B$ are given by

$$d_{fk} = I_{fk} \alpha_{fk} / \pi_k^f, \tag{12}$$

where I_{fk} is an indicator variable for element k of frame f, α_{fk} is the conditional sample membership indicator for element k in frame f, and $\pi_k^f = E(\alpha_{fk} \mid I_{fk} = 1)$ is the conditional probability of the selection of element k from frame f.

Considering an agricultural survey with Frame A as an area frame and Frame B as a list frame, Hartley's estimator can be applied immediately in surveys where Frame A is of Type 3 and Frame B is of Type 1, as described in Table 1 above. In this situation, elements (such as holders' addresses) can be selected directly using one-stage probability sample designs, such as simple random sampling, systematic sampling or stratified sampling. If at least one of the frames has clusters as elements (Types 2, 4, 5, or 6), a multiple-stage sampling design is necessary to identify the elements and then make it feasible to match the sampled units (the identifiability requirement).

One advantage of the estimator shown in Equation 15 below is the possibility of simplifying the estimation process by choosing zero as the value of ψ. In this case, the resulting estimator is called a screening estimator, because to be effective, the procedure requires screening and eliminating, from Frame A, all component units that also belong to Frame B.

Supposing that a simple random sampling is applied to each frame, and denoting the domain population variances by σ_a^2, σ_b^2 and σ_{ab}^2, the variance (see Equation 14 below) is approximated by the variance for stratified samples with an allocation proportional to the domain sizes (ignoring finite population correction factors):

$$Var(\hat{t}_H) = \frac{N_A^2}{n_A}\left[\sigma_a^2\left(1-\frac{N_{ab}}{N_A}\right)+\phi^2\sigma_{ab}^2\frac{N_{ab}}{N_A}\right]+\frac{N_B^2}{n_B}\left[\sigma_b^2(1-\frac{N_{ab}}{N_B})+(1-\phi)^2\sigma_{ab}^2\frac{N_{ab}}{N_B}\right]. \quad (13)$$

5.1.2. The Fuller- Burmeister Estimator

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

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

Fuller and Burmeister (1972) proposed an estimator that uses sample information from the frames to estimate N_{ab}. Let

$$\hat{N}_{ab}^A = \sum_{k\in S_{ab}^A}\frac{1}{\pi_k^A}, \qquad \hat{N}_{ab}^B = \sum_{k\in S_{ab}^B}\frac{1}{\pi_k^B} \quad (14)$$

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

$$\hat{t}_{FB} = \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \sum_{k \in S_b} \frac{y_k}{\pi_k^B} + \sum_{k \in S_{ab}} y_k^* + p_2\left(\hat{N}_{ab}^A - \hat{N}_{ab}^B\right), \tag{15}$$

where

$$y_k^* = \begin{cases} p_1 \dfrac{y_k}{\pi_k^A}, & if \quad k \in S_{ab}^A \\ (1-p_1)\dfrac{y_k}{\pi_k^B}, & if \quad k \in S_{ab}^B \end{cases} \tag{16}$$

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

The Fuller-Burmeister estimator written in multiplicity notation is presented next. The estimators for $\hat{N}_{ab}$ based on Frames A and B are given by

$$\hat{N}_{ab}^f = \sum_k d_{fk} \psi_k^f, \qquad f = A, B. \tag{17}$$

where $\psi_k^f = 1$ if element k is in both frames, and $k \in S_{ab}^A$ and $k \in S_{ab}^B$ respectively.

Then, the Fuller-Burmeister estimator is

$$\hat{t}_{FB} = \sum_f \sum_k d_{fk} y_k \phi_k^f + \psi_B (\hat{N}_{ab}^A - \hat{N}_{ab}^B) = \sum_f \hat{Y}_f^* + \psi_B (\hat{N}_{ab}^A - \hat{N}_{ab}^B), \quad (18)$$

where $\hat{Y}_f^* = \sum_k d_{fk} y_{fk}^*$ and $y_{fk}^* = y_k \phi_k^f$. In this expression, ϕ is a weighting constant ($0 \le \phi \le 1$) chosen to minimize the variance of the estimator $\hat{t}_{FB}$, such that $\phi_k^A = \psi_A$ and $\phi_k^B = 1 - \psi_A$ if element k is in both frames, $k \in S_{ab}^A$ and $k \in S_{ab}^B$ respectively, and $\phi_k^f = 1$ if element k is in only one frame.

The weights for each frame, $f = A, B$, follow Equation 16.

In agricultural surveys, the Fuller-Burmeister estimator can be readily applied and accommodate complex sampling designs in each frame, if necessary.

5.1.3. Single Frame Type Estimator

Bankier (1986) and Kalton and Anderson (1986) proposed a single frame type estimator that relies on a set of sampling weights which enable the estimator to be written as the sum of only two Horvitz-Thompson estimators, each covering the sample data from one of the frames. The single frame type estimator can also be written in general form (12) as shown below:

$$\hat{t}_B = \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \sum_{k \in S_b} \frac{y_k}{\pi_k^B} + \sum_{k \in S_{ab}} \frac{y_k}{\pi_k^A + \pi_k^B} \quad (19)$$

Using multiplicity notation, the estimator is as follows:

$$\hat{t}_B = \sum_f \sum_k d_{fk} y_k \phi_k^f = \sum_f \hat{Y}_f^*, \quad (20)$$

where $\hat{Y}_f^* = \sum_k d_{fk} y_{fk}^*$ and $y_{fk}^* = y_k \phi_k^f$. In this expression, ϕ is such that $\phi_k^f = \frac{\pi_k^f}{\left(\pi_k^A + \pi_k^B\right)}$ if element k is in both frames, $k \in S_{ab}$, and $\phi_k^f = 1$ if element k is in only one frame. The weights for each frame, $f = A, B$, follow Equation 16.

Originally, Bankier discussed the scenario in which stratified probability samples were selected from Frames A and B. Considering Frame A as an area frame, the stratification of segments by land use is common in agricultural surveys. Stratification can also be applied to list frame B when auxiliary information is available. Applying the single frame type estimator in this case has the advantage of producing estimates that are based on a simple formula which combines estimates obtained separately from Frames A and B.

5.1.4. The Skinner- Rao 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

$$\hat{t}_{SR} = \frac{N_A - \hat{N}_{ab,SR}}{N} \sum_{k \in S_a} \frac{y_k}{\pi_k^A} + \frac{N_B - \hat{N}_{ab,SR}}{N} \sum_{k \in S_b} \frac{y_k}{\pi_k^B} + \sum_{k \in S_{ab}} \gamma y_k^*, \tag{21}$$

where

$$y_k^* = \begin{cases} \frac{y_k}{\pi_k^A}, & if \quad k \in S_{ab}^A \\ \frac{y_k}{\pi_k^B}, & if \quad k \in S_{ab}^B \end{cases} \qquad \gamma = \frac{\hat{N}_{ab,SR}}{\left(\pi_k^A \hat{N}_{ab}^A + \pi_k^B \hat{N}_{ab}^B\right)}. \tag{22}$$

$N_{ab,SR}$ 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 $\alpha_3 = n_A \hat{N}_{ab}^A N_B + n_B \hat{N}_{ab}^B N_A$.

Using multiplicity notation, the Skinner-Rao estimator assumes the form

$$\hat{t}_{SR} = \sum_f \sum_k w_k^f y_k = \sum_f \hat{Y}_f^*, \tag{23}$$

where

$$w_k^f = \begin{cases} d_{fk}, & if \quad k \in S_{dom} \\ \phi_k^f d_{fk}, & if \quad k \in S_{ab}^f \end{cases}. \tag{24}$$

In this expression, $\hat{Y}_f^* = \sum_k w_k^f y_k$, $dom = \{a, b\}$ and $\phi_k^f = \dfrac{\hat{N}_{ab,SR}}{\left(\pi_k^A \hat{N}_{ab}^A + \pi_k^B \hat{N}_{ab}^B\right)}$ if element k is in both frames, $k \in S_{ab}^A$ and $k \in S_{ab}^B$ respectively, and $\phi_k^f = 1$ if element k is in only one frame. The weights for each frame, $f = A, B$, are given by Equation 16.

The Skinner-Rao estimator can efficiently accommodate complex sampling designs.

Given a set of two or more agricultural frames described in Table 1 above, and a dual frame sampling design that incorporates complex features such as stratification and multiple stages, situations may arise in which more than one estimator can be feasibly applied. In these cases, running brief Monte Carlo simulation experiments based on information from available past surveys may be useful in guiding the choice. The choice of estimator should take into consideration not only statistical performance but also simplicity. Therefore, screening-type estimators are always sound candidates and should not be discarded.

Although dual frame methods have a great potential to generate highly precise agricultural statistics, they should be adopted only pursuant to an evaluation of the costs and benefits in terms of precision and accuracy versus the greater complexity due to working with two frames, as compared to a single frame survey option. Vogel (1975) describes the operational challenges of using a dual frame approach in agricultural settings. Lohr and Rao (2000) also note 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 questionnaires. In these cases, bias may be introduced into the estimates. Moreover, since the dual frame estimators rely on the identifiability requirement, the estimates may be biased if problems with classifying sampled units in the correct domains arise.

Lohr and Rao (2000) provide a description of resampling methods for the variance estimation of dual frame estimators. Mecatti (2007) shows that the multiplicity estimator has a closed variance form, which is an advantage in terms of simplicity.

6

Dual frame sample size allocation

Demnati et al. (2007) outline several strategies for determining optimal frame sample size allocation when estimating the population total of one or more characteristics of interest under a dual-frame survey approach. In this Section, examples of sample size allocation that minimize costs, subject to constraints on the variances of dual frame estimators for population totals, are described for Hartley's dual frame estimator and Hartley's dual frame estimator by calibrating on sizes N_A and N_B .

In a dual frame survey, it is assumed that frames A and B together cover the population of interest U. Demnati et al. also assume that $\mathsf{A} = \mathsf{U}$ is a complete frame and that $\mathsf{B} \subset \mathsf{U}$ is incomplete, and that A is expensive to sample and B is cheap to frame. They consider the Frames A and B having sizes of $N_A = N$ and $N_B < N$ respectively. First, the estimation of the population total of one characteristic of interest is illustrated; second, the population total of more than one characteristic of interest is estimated. In dual frame surveys, the population total t of a characteristic of interest y can be expressed as

$$t = \sum_f \sum_k I_{fk} y_k \phi_k^f \,,\; f = A, B. \tag{25}$$

Here, I_{fk} is a indicator variable for element k of frame f, $\phi_k^A = \psi$ and $\phi_k^B = 1 - \psi$ if element k is in both frames, with $0 \le \psi \le 1$, and $\phi_k^f = 1$ if element k is only in one frame. The weights for each frame $f = A, B$ are given by Equation 15 above.

6.1. Sample Allocation with Hartley's Estimator

Hartley's dual frame unbiased estimator of total t is given as in Equation 16 above:

$$\hat{t}_H = \sum_f \sum_k d_{kf} y_{fk} \phi_k^f = \sum_f \hat{Y}_f^*, \tag{26}$$

where $\hat{Y}_f^* = \sum_k d_{fk} y_{fk}^*$ and $y_{fk}^* = y_k \phi_{fk}$.

Assuming that independent samples are drawn from each frame by means of simple random sampling (SRS), with sample size n_f, $f = A, B$, the sampling variance of is given by

$$Var(\hat{t}_H) = Var(\sum_f \hat{Y}_f^*) = \sum_f Var(\hat{Y}_f^*), \tag{27}$$

where

$$Var(\hat{Y}_f^*) = N_f^2 (1 - n_f / N_f) S^2(y_f^*) / n_f = N_f^2 S^2(y_f^*) / n_f - N_f S^2(y_f^*),$$

$$S^2(y_f^*) = \sum_f I_{fk} (y_{fk}^* - \bar{y}_f^*)^2 / (N_f - 1), \text{ and}$$

$$\bar{y}_f^* = \sum_f I_{fk} y_{fk}^* / N_f .$$

Considering p characteristics of interest $y_1,\ldots,y_p$, the population total Y_j of a characteristic of interest y_j and its correspondent sampling variance can be expressed as

$$\hat{t}_{Hj} = \sum_f \sum_k d_{fkj} y_{kj} \phi_k^f = \sum_f \hat{Y}_{fj}^*, \tag{28}$$

where $\hat{Y}_{fj}^* = \sum_k d_{fkj} y_{fkj}^*$ and $y_{fkj}^* = y_{kj}\phi_k^f$, and

$$Var(\hat{t}_{Hj}) = -\sum_k N_{fj} S^2(y_{fj}^*) + \sum_k N_{fj}^2 S^2(y_{fj}^*)/n_f // \nu_{0j} + \sum_k \nu_{fj}/n_f, \tag{29}$$

where $\nu_{0j} = -\sum_k N_{fj} S^2(y_{fj}^*) and \nu_{fj} = N_{fj}^2 S^2(y_{fj}^*)$. To determine the optimal values of n_A and n_B for a specified ϕ, the cost function $C = c_0 + \sum_k c_j n_j$ was minimized, subject to constraints on the p variances:

$$Var(\hat{t}_{Hj}) \le V_j, j = 1,..., j,$$

where V_j are specified tolerances, considered as $V_j = (\delta_j Y_j)^2$.

6.2. Simulation Data

The illustration of results is based on artificial data, built as follows:

1. A population is generated with four characteristics $(y_{1k}, y_{2k}, y_{3k}, y_{4k})$ of size $N = 1000$, such that

$$y_{1k} = B(1,0.6),\ y_{2k} + 50 + 16 \times \varepsilon_k,$$

where $\varepsilon_k \ N(0.1)$, $y_{3k} \ B(1, p_k)$ and
$p_k = exp(0.1+1\times J_{2k})/(1+exp(0.1+1\times J_{2k}))$, where $p_k \approx 0.75$ for $J_{2k} = 1$, and $p_k \approx 0.52$ for $J_{2k} = 0$.

2. The frame A was set to be complete with membership indicator $J_{1k}. = 1$ The membership indicator for frame B was generated from $J_{2k} \ B(1,0.6)$ with 60% coverage.

3. SRSs were taken from each frame.

6.3. Simulation Results

Demnati et al. (2007) set the cost values $c_0 = 0$, $c_1 = 1$ and $c_2 = 0.5$ and $c_2 = 0.2$. The tolerances are set as $\delta_j = 0.05$ for $j = 1,\ldots,4$. Also, ϕ is set as 0.5 and as the optimal value. To determine the optimal value of ϕ, different values of ϕ between 0 and 1 were used for a repeated optimization process. For $c_2 = 0.5$, the optimal result $\phi = 0.87$ was found, and for $c_2 = 0.2$, $\phi = 0.64$. Comparing the results in Table 2 below, it can be seen that C_{min} for cost $c_2 = 0.2$ is lower compared to the case in which sampling is carried out only from the complete frame A, and that when the optimal value of ϕ is used, that measure is lower.

Table 2 – Optimal n_1 n_2 and C_{min} : Hartley's estimators

	$\psi 7$	n_1	n_2	C_{min}
Complete Frame A	0.5	203		203
C_2 = 0.05	0.5	174	92	220
C_2 = 0.02	0.5	161	134	188
C_2 = 0.05	.87	190	23	202
C_2 = 0.02	.64	166	96	186

7

Auxiliary data in a dual frame survey

In any survey, auxiliary information taken into account either at the design or at the estimation stage may lead to an increase in the estimate's precision. In dual frame agricultural surveys, auxiliary information may be available from past agricultural censuses, from certain types of list frames, or even from the area measure of segments defined in the area frame. In this Section, we discuss two alternative ways to use auxiliary information. The first is use of a multivariate probability-proportional-to-size sampling (MPPS) scheme. The second is the definition of a ratio-type estimator.

7.1. Multivariate Probability-Proportional-to-Size Sampling

Sampling with MPPS extends PPS schemes, in the sense that it uses more than one category of auxiliary information to compose the inclusion probability. The efficiency of PPS sampling depends on the degree of correlation between the variable of interest and auxiliary variables. As long as this correlation is strong, the Horvitz-Thompson (1952) estimator is the more efficient choice, rather than an equal-probabilities design. However, if the correlation is weak or null, the HT estimator may not be the best choice: indeed, although it would still be unbiased, its variance would be very large. The multipurpose characteristic of agricultural surveys would make it inefficient to adopt a simple PPS sampling, because this would imply high variances for a series of variables of interest without any correlation with the auxiliary information used to generate the inclusion probability. The MPPS is capable of providing a compromise solution to the problem. Consider $K \geq 2$ variables of interest (items) of a multipurpose survey, each with at least one auxiliary (size) variable available:

$$\begin{cases} y_{i(1)} & x_{i(1)} \\ \vdots & \vdots \\ y_{i(k)} & x_{i(k)} \\ \vdots & \vdots \\ y_{i(K)} & x_{i(K)} \end{cases} \tag{30}$$

$\forall i \in U$.

The same sample is used for estimation purposes on every item, so that a unique π_i must be defined for each population unit. In an MPPS sampling,

$$\text{MPPS} \overset{def}{\rightarrow} \pi_i = f\left(n_k \frac{x_{i(k)}}{X_k}, \quad k = 1 \cdots K \right), \tag{31}$$

where

- f is a function to be chosen
- $X_k = \sum_{i \in U} x_{i(k)}$ is the auxiliary (size) population total and
- n_k is the number of units to be selected for the k-th item.

In the examples of the China census (Yong et al. 2006) and NASS (Hicks et al. 1996), it is suggested to use

$$\pi_i = \min\left\{ 1, \quad \max\left(n_k \frac{x_{i(k)}}{X_k}, \quad k = 1 \cdots K \right) \right\} \tag{32}$$

7.2. Ratio - Type Estimators

Consider the estimation of a population total (or mean) incorporating available auxiliary information from one of the frames at the estimation stage. Using Hartley's dual frame approach, Ferraz and Coelho (2007) investigate ratio-type dual frame estimators written in the general form:

$$\hat{t}_{y,r} = \hat{t}_{ya,r} + \hat{t}_{yb,r} + \hat{t}_{yab,r}, \tag{33}$$

where $\hat{t}_{ya,r}, \hat{t}_{yb,r}$, and $\hat{t}_{yab,r}$ are ratio-type estimators for population totals in domains a, b and ab, respectively. It is also possible to modify the GMHT estimator to generate ratio-type estimators. Let Y and X denote the response variable and its auxiliary variable, respectively. Thus, the GMHT estimator can be modified for the scenario of a ratio-type estimator in two ways, called Generalized Multiplicity Ratio Estimators 1 and 2 (GMRE1 and GMRE2), respectively. The first form considers the multiplicity estimator applied for each variable, generating a ratio-type multiplicity estimator as follows. Let $t_{x(q)}$ denote the known value of a population total of auxiliary information in each frame. Thus,

$$\hat{t}_{y,GMRE1} = \left(\frac{\sum_{q=1}^{Q} \sum_{k \in U_q} y_k \phi_{q(k)} \dfrac{\delta_{k(q)}}{E\left(\delta_{k(q)}\right)}}{\sum_{q=1}^{Q} \sum_{k \in U_q} x_k \phi_{q(k)} \dfrac{\delta_{k(q)}}{E\left(\delta_{k(q)}\right)}} \right) t_{x(q)}. \tag{34}$$

The second form considers the use of ratio-type estimators modifying the original form of the GMHT estimator. We have

$$\hat{t}_{y,GMRE2} = \sum_{q=1}^{Q} \left\{ \left(\frac{\sum_{k \in U_q} y_k}{\sum_{k \in U_q} x_k} \right) \phi_{q(k)} \frac{\delta_{k(q)}}{E\left(\delta_{k(q)}\right)} \right\} t_{x(q)}. \tag{35}$$

8

Conclusion

This Technical Report explored the subject of linking frames for agricultural surveys. A review of the literature concerning methods for linking frames at the design and estimation stage of a survey was provided, emphasizing the potential for using a dual frame approach as an efficient way to take advantage of information from area frames and list frames.

Linking information from different registers is an essential task for building list frames with full coverage. The increasing computational ability to handle massive data sets concretizes the possibility of exploring administrative data as one of these source registers. However, this approach should be taken only if the different sources contribute essential information to the frame and if the record matching yields extremely reliable results. The cost of this 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. It is recommended that studies assessing these costs be conducted.

An alternative to dealing with a list frame building process is the use of available 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 to avoid bias. Since using multiple frames entails a significant increase in complexity, the design of a dual frame survey using an area and a list frame may be a good compromise. In this case, using an area frame has the further advantage of enabling the objective assessment of land characteristics, such as cultivated area, if necessary.

The choice of the sample design applied to each frame in a dual frame survey depends on the types of frames available (see Table 1 above). The frames with clustered elements require a multiple-stage design for the identifiability requirement to be met. In addition, more than one dual frame estimator may be applied. In these cases, it is recommended to run Monte Carlo simulation experiments based on available agricultural information to guide the choice. In these simulations, the estimators should also consider their respective screening versions.

Further improvements in the statistical efficiency of dual frame estimators may be obtained by studying ways to incorporate auxiliary information into their functional form. In this regard, MPPS sampling and ratio-type estimators for dual frame surveys can be of assistance.

References

Abaye, A.T. 2013. Master sampling frames for agricultural and rural statistics in Ethiopia. In *Proceedings of the Sixth International Conference on Agricultural Statistics* (pp. 52-54). IBGE Publication: Rio de Janeiro, Brazil,.

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.

Benedetti, R., Bee, M., Espa, G. & Piersimoni, F. 2010. *Sampling Methods for Agricultural Surveys*. Wiley: Chichester, UK.

Carfagna, E. 2001. *Multiple frame sample surveys: advantages, disadvantages and requirements*. In *International Statistical Institute, Proceedings – Actes – Invited papers Seoul (2001)* (pp. 253-270), ISI Publication: The Hague, Netherlands. 2004. *List frames, area frames and administrative data, are they complementary or in competition?*, Paper prepared for the Third International Conference on Agricultural Statistics. 2-4 November 2004. Cancún, Mexico.2007. A comparison of area frame sample designs for agricultural statistics. In *Bulletin of the International Statistical Institute Vol. LXII: 56th Session Proceedings* (pp. 2143-2146), Special Topics Contributed Paper Meetings (STCPM11). ISI Publication: The Hague, Netherlands.

Carfagna E. & Carfagna A. 2010. Alternative sampling frames and administrative data; which is the best data source for agricultural statistics?, in Benedetti R., Bee M., Espa G., Piersimoni F. (eds) *Agricultural Survey Methods*. Wiley: Chichester, UK.

Demnati, A., Rao, J.N.K., Hidiroglou, M.A. & Tambay, J.-L. 2007. On the allocation and estimation for dual frame survey data. In *Proceedings of the Survey Research Methods Section, American Statistical Association* (pp. 2938-2945). ASA Publication: Alexandria, VA, USA.

D'Orazio, M., Di Zio, M. & Scanu, M. 2006. *Statistical Matching: Theory and Practice*. Wiley: Chichester, UK.

FAO. 1989. *Sampling Methods for Agricultural Surveys*, FAO Statistical Development Series No. 3. FAO Publication, Rome. 1996. *Multiple Frame*

Agricultural Surveys, FAO Statistical Development Series No. 7, Volume 1. FAO Publication, Rome. 1998. *Multiple Frame Agricultural Surveys*, FAO Statistical Development Series No. 10, Volume 2. FAO Publication, Rome.

Fellegi, I.P. & Sunter, A.B. 1969. A Theory for Record Linkage. *Journal of the American Statistical Association*, 40: 1183-1210.

Ferraz, C. & Coelho, H.F.C. 2007. Ratio Type Estimators for Stratified Dual Frame Surveys. In *Bulletin of the International Statistical Institute Vol. LXII: 56th Session Proceedings* (pp. 4936-4939). ISI Publication: The Hague, Netherlands.

Fuller, W.A. & Burmeister, L.F. 1972. Estimators for samples selected from two overlapping frames. In *Proceedings of the Social Statistics Section, American Statistical Association* (pp. 245-249), ASA Publication: Alexandria, VA, USA.

Gallego, F.J. 1995. Sampling frames of square segments. Report EUR 16317. Office for Publications of the European Communities: Luxembourg. 2013. The use of a point sample as a master frame for agricultural statistics, Paper prepared for the Sixth International Conference on Agricultural Statistics, 23-25 October 2013. Rio de Janeiro, Instituto Brasileiro de Geografia e Estatística (IBGE).

Gallego F.J., Delincé J. & Carfagna E. 1994. Two-Stage Area Frame Sampling on Square Segments for Farm Surveys. *Survey Methodology*, 20(2): 107-115.

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.

Hartley, H.O. 1962. Multiple frame surveys. In *Proceedings of the Social Statistics Section, American Statistical Association* (pp. 203-206), ASA Publication: Alexandria, VA, USA.

Horvitz, D.G. & Thompson, D.J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47: 663-685.

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

Lavallée, P. 2005. Quality indicators when combining survey data and administrative data, In *Proceedings of Statistics Canada Symposium 2005: Methodological Challenges for Future Information Needs*. Available at http://www.statcan.gc.ca/pub/11-522-x/11-522-x2005001-eng.htm.

Lohr, S. & Rao, J.N.K. 2000. Inference from dual frame surveys. *Journal of the American Statistical Association*, 95(449): 271-280.

Mecatti, F. 2007. A Single Frame Multiplicity Estimator for Multiple Frame Surveys. Survey *Methodology*, 33: 151-158.

Mecatti, F. & Singh, A.C. 2014. Estimation in Multiple Frame Surveys: A Simplified and Unified Review using Multiplicity Approach. *Journal de la Societé Française de Statistique*, 155(4): 51-69.

Santos, D., Freitas, M., Lila, M., Arantes, S., Dantas, T. & Santos, V. 2013. Brazilian agricultural survey system: a description of sampling methods. In *Proceedings of The Sixth International Conference on Agricultural Statistics* (pp. 251-258). IBGE Publication: Rio de Janeiro, Brazil.

Särndal, E., Swensson, B. & Wretman, J. 1992. *Model assisted sample surveys*. Springer Series in Statistics, Springer-Verlag: New York, USA.

Shin, H., Molinari, N. & Wolter, K. 2008. A dual-frame design for the National Immunization Survey. In *JSM Proceedings* (pp. 4368-4375), Survey Research Methods Section. ASA Publication: Alexandria, VA, USA.

Selander, R., Svensson, J., Wallgren, A. & Wallgren, B. 1998. *How should we use IACS data?* Statistics Sweden Publication: Stockholm.

Skinner, C. & Rao, J.N.K. 1996. Estimation in dual frame surveys with complex designs. *Proceedings of the Survey Methods Section, American Statistical Association* (pp. 63-68). ASA Publication: Alexandria, VA, USA.

Vogel, F.A. 1975. Surveys with overlapping frames – problems in applications. In *Proceedings of the Social Statistics Section, American Statistical Association* (pp. 694-699). ASA Publication: Alexandria, VA, USA.

Wallgren, A. & Wallgren, B. 1999. *How can we use multiple administrative sources?* Statistics Sweden Publication: Stockholm.

Winkler, W.E. & Thibaudeau, Y. 1987. *An Application Of The Fellegi-Sunter Model Of Record Linkage To The 1990 U.S. Decennial Census*. In U.S. Decennial Census Tech-nical Report. US Bureau of the Census Publication: Washington, D.C.

Yitzkov, T. & Azaria, H. 2003. Record Linkage in an Integrated Census. In *Proceedings*, FCSM 2003 Research Conference, 17-19 November 2003, Washington, D.C.

Global Strategy to Improve Agricultural and Rural Statistics