Updating sampling frames for agricultural statistics: approaches, challenges and issues

Elisabetta Carfagna*
University of Bologna, Bologna, Italy – elisabetta.carfagna@unibo.it

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

Abstract

The traditional approach for producing agricultural statistics adopted in most developed countries is the following (see Benedetti et al. eds. 2010): a complete enumeration census is carried out every 5-10 years. The census is carried out by interviewing the farmers and allows generating the list frame that is used for all kinds of sample surveys of farms.
In this paper, the quality of data collected by a census and the coverage of the list generated by a census are discussed. The impact of incomplete or out of date sampling frames for agricultural statistics is analysed. Some approaches for updating the sampling frames, the problems involved and the difficulties faced are addressed. Then, different kinds of sampling frame and estimation approaches are analysed, taking into consideration the specificities of countries.

Keywords: agricultural statistics, updating list frames, area and multiple frames.

1. Introduction

The traditional approach for producing agricultural statistics adopted in most developed countries is the following (see Benedetti *et al*. eds. 2010): a complete enumeration census is carried out every 5-10 years. The census is carried out by interviewing the farmers and allows generating the list frame that is used for all kinds of sample surveys of farms; thus, it is a master sampling frame for agricultural statistics. Data are collected through mail, email, personal interviews, computer assisted personal interviews, computer assisted telephone interviews, or the web. Before using the census list as a master sampling frame for annual sample surveys, the following crucial questions should be answered: Which is the quality of the data collected by the census? Is the census list complete and without duplications? When will it become out of date? Which is the impact of incompleteness and obsolescence? How can the list be updated? Is an alternative approach feasible? In which cases should it be preferred?
In this paper, we address these questions, focusing on the impact of out of date sampling frames for agricultural statistics and the issues related to some approaches for updating the sampling frames, the problems involved and the difficulties faced, taking into consideration the specificities of countries.

2. Quality of data collected by the census

For a discussion on the quality of census data, we take advantage of a very accurate assessment made on the results of the complete enumeration census of agriculture carried out in Italy in 2010 (Mazziotta, 2013). At the end of the census data collection, a sample survey for assessing the quality of collected data was designed. A stratified random sample of about 50,000 farms was selected and the farmers were interviewed through computer assisted telephone interviews in the period from 20 May 2011 to January 2013. At a country level, the following results were obtained: a bias per cent of 0.6 for wheat, -2.9 for corn and -3.0 for set-aside. They confirm that the data collected through the census of agriculture are not very reliable for area of crops. Besides that, the following biases per cent, at country level, were obtained: -0.2 for arable land, -1.4 for permanent wood crops, -2.9 for permanent grassland and utilized pasture, -1.3 for utilized agricultural area, -2.9 for total area and -2.8 for the irrigated area; worse biases were obtained for livestock.
These results show that the complete enumeration census systematically underestimates the main structural variables, which are generally used for stratification, when annual sample surveys are designed. In addition, the level of the bias varies in the different regions of the country.

3. Coverage of the list of farms generated by a census of agriculture

For the Italian agricultural census, the coverage was assessed on the basis of an area sample. Around 1,500 sheets of cadastral maps (areal units in which each municipality is subdivided – secondary sampling units) were selected from a sample of municipalities (primary sampling units).
The owners of the parcels in the selected sheets of cadastral maps were identified, on the basis of the cadastral archive, and interviewed. 21,588 farmers were interviewed (1.620.884 active farms and 34.070 temporary inactive farms were identified by the agricultural census).
Estimates were computed in the framework of the indirect sampling (Lavallée, 2007), and the weights (Lavallée and Rivest, 2012) were assigned based on the selection probability of each sheet of cadastral map and the number of sheets in which a farm has parcels (derived from the interview).
A sophisticated record linkage procedure was implemented in three successive steps: deterministic, probabilistic and manual, involving various kinds of administrative registers. 81.4 % of farms in the area frame were included in the census list; 5.2 % of farms in the area frame were present in the census list with different characteristics, 1.7 % of the farms in the area frame had multiple links with census list, and 11.7 % of the farms in the area frame had no link with the census list. Of course, the percentage of farms in the area and in the census list decreases for small farms: 71% and 78.2 % for farms with utilized agricultural area in the range (0.01 - 0.99 hectares) and in the range (1 -1.99 hectares) respectively. This level of coverage is in line with most developed countries.
These results of the quality assessment of the census data stimulates a reflection, if the main aims of the agricultural census are creating the list of all farms (including small ones) to be used as master sampling frame, with accurate structural information for stratification and producing estimates for very small administrative domains, at least once every 5-10 years.

4. Coverage of the census-sampling frame for specific categories of farms

List frames generated by a census can have a very low coverage or can deteriorate very rapidly for specific categories of farms, even after the integration of the census list with some administrative registers.
Let us show some results of a study conducted in Campania Region, in Italy, in 2002, just two years after the 5th census of agriculture (Giovacchini 2012). An area frame sample survey of farms cultivating flowers was conducted and the comparison was done with the census list updated with registers, like the register of farmers for the use of pesticides. The subsidies register was not used, since this kind of farms does not receive subsidies. The under-coverage, came out to be 48%; 54% if only farms with a surface smaller than or equal to half an hectare is taken into account. Note that farms of this size account for 74% of farms cultivating flowers detected by the area sample survey. Moreover, in this study, farms were selected through a grid of points located on the selected square segments; this means that farms were selected with probability proportional to size, thus small farms had lower probability than large ones to be included in the area frame sample.

5. Updating the census list with administrative data

Various kinds of administrative registers are generally used for updating the census list. The quality of the result depends on the administrative data that can be used and on the consistency of the identifiers of the units in the different registers. The over and under coverage can be high even if good administrative data, very sophisticated record linkage procedures and geo-location of administrative information are used, as showed by the following experiment. Several kinds of administrative data were taken into consideration for updating the Italian census list in 2008 (8 years after the census). Main registers used were the lists related to farms that apply for subsidies, livestock farms, agrarian income, cadastre, taxes, social security and specific lists created by regional authorities. A sample of 15,682 units was selected out of a subset of 80 municipalities. Enumerators used a web-based data collection system developed on purpose, in order to ensure accurate data collection. The result was that only 39.15% of the farms included in the integrated list were considered existing and active by the test. 44.74% of the

farms in the integrated list were not active and 16.11% of them were not identified through the test (Berntsen and Viviano, 2011).
This level of over-coverage implies that, if such a list is used for a sample survey, the enumerators waste much time trying to identify farmers, which then prove to be inactive. Moreover, distinguishing inactive farmers from total non-responses is difficult. Finally, the risk of producing biased estimates is high, unless an accurate estimate of the over-coverage is available. These considerations suggests adopting this approach only where the reliability of administrative data used for updating the census list is very high and the definitions adopted by administrative registers are compatible with the ones of the census.

6. Linking the Population and Housing Census

Now we address the question: Which approach could represent an alternative to a census of agriculture? A recent proposal comes from FAO and UNFPA (2012), for avoiding facing the cost of the agricultural census. It has been adopted in some countries, like Mozambique and Burkina Faso.
A list frame is created based on the population census, the list of farms or agricultural households identified on the basis of specific agricultural questions included in the population census questionnaire. This approach is promising for countries where agriculture is not an important economic sector, like small islands. More work is needed for testing the quality of data collected using long questionnaires and the coverage of the list of farms generated from the population census; particularly, the entity of under and over coverage in different categories of countries should be assessed. Finally, we have to remark that the list frame of farms generated through the module on agriculture submitted to the households presents very few auxiliary variables; thus, the efficiency of the sample designs for annual sample surveys is very low, and this may have a strong impact on annual survey costs. For more details and an analysis of advantages, disadvantages and requirements see Keita and Gennari (2013) and Carfagna *et at*. (2013).

7 Linking frames at the survey design stage

Combining information matching records from several sources at the survey design stage is a way to build and update a sampling frame. The goal of linking area and list frames at the frame component unit level is a challenge as it depends on existing a linking connection between area units and list units, before the survey field work. In some countries, population and agricultural censuses have generated a database with point locations that may provide such link.
The problem of matching records from two frames leads to the subject of record linkage. Several authors have been contributing to the development of a theory about the problem of record matching. Fellegi and Sunter (1969) wrote one of the first papers on this topic. In the search for a decision rule that establishes 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. Fellegi and Sunter established conditions under which a linkage rule is optimal.
Winkler and Thibaudeau (1987) described an application of the Fellegi-Sunter model to the 1990 US decennial population census. Yitzkov and Azaria (2003) application of record linkage theory to the Israel's 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.

8. Creating master sampling frames for agricultural statistics through integration of registers

Several North European countries have been using registers more and more extensively, in order to reduce the cost and the respondent burden due to data collection (see for example Wallgren and Wallgren, 2010). Subsidies are an important source of data in European countries; however, their use for direct tabulation is not feasible, as explained in Carfagna and Carfagna, 2010. In Sweden, a country that initiated to make an extensive use of registers for statistics decades ago, the annual agricultural statistics are produced through sample surveys, based on a list frame built through registers, mainly tax files. Creating a master sampling frame by integrating different kinds of registers is not an easy task, consider that, in 2009, the business register and the farm register at Statistics Sweden were not harmonized yet (Wallgren and Wallgren, 2010). Moreover, the risk of under-coverage could be high,

since units below a threshold required to be registered or pay taxes are generally excluded, as well as the units that do not apply for subsidies. The kind of list frame generally include commercial farms, but are not likely to include small-scale farms and subsistence farming units (see Carfagna and Carfagna, 2010 and Carfagna *et al*. 2013).

9 Multiple frames for agricultural surveys

Examples of countries working on a composition of several frames into a master frame for integrating agricultural and population surveys include Brazil and Ethiopia. In the Sixth International Conference on Agricultural Statistics, statisticians from both countries presented a brief talk about the subject.

The Central Statistical Agency of Ethiopia (CSA) is responsible for carrying out the agriculture surveys in the country. They collect data 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, furnishing a frame list of households to be selected as secondary sampling units. In order to improve agricultural data quality, CSA is currently studying the possibility of adopting the following composition between an area and a list frame: an area frame sampling is conducted with EA as primary sampling units, and segments of size 40 hectare as secondary sampling units, stratified by land cover classification. In addition, data from commercial farms is also collected, using a list frame. Information from area and list frames should be combined to improve estimates.

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 frame and list frame aiming to provide coverage to the target population of agricultural establishments in the country. Accordingly with Santos et al. (2013), "*Establishments 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 into a master frame for agricultural surveys. Linking information from different sources can be done either at the design or at the estimation stage of a survey. The strategy of trying to link data from several frames at the design stage of a survey corresponds to concentrate efforts to build a single frame from multiple sources, and carry out a single frame survey. One of the problems faced when doing this is to appropriately match records among different registers, what leads to the subject of record linkage. However, in addition to the challenges posed when no perfect matching are detected, the same problems of frame maintenance and costs would apply to the survey resulting from such approach.

The strategy of linking information from different frames at the estimation stage, on the other hand, deals with the idea of combining information from each frame with no need for matching records before the survey data collection process. Such a strategy offers flexibility to design a survey sampling from each frame independently, and it is less error prone than the strategy of linking data at the design stage. The multiplicity estimator, proposed by Mecatti (2007) provides 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 whose union is assumed to cover the target population $\mathsf{U} = \bigcup_{q=1}^{Q} U_q$. The frames can overlap themselves and some can even furnish full population coverage. In this approach, independent samples denoted by $S_1, S_2, \ldots, S_Q$ are selected from each one of the Q frames with no need to keep the same probability sample design for them. Consider the goal of estimating the population total of a study variable,

$$t = \sum_{k \in \mathsf{U}} y_k \quad , \tag{4}$$

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

$$t = \sum_{k \in U} y_k = \sum_{q=1}^{Q} \sum_{k \in U_q} y_k \phi_{q(k)}, \tag{5}$$

where $0 \le \phi_{q(k)} \le 1$, in general but not necessarily, and $\sum_{q=1} \phi_{q(k)} = 1$ are the multiplicity adjustment factors corresponding to q th frame and the k th 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

$$\hat{t} = \sum_{q=1}^{Q} \sum_{k \in U_q} y_k \phi_{q(k)} \frac{\delta_{k(q)}}{E(\delta_{k(q)})}. \tag{6}$$

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

10. Area and multiple frame sample surveys

The estimator proposed by Mecatti and Singh (2014) has the crucial requirement that the multiplicity of each sample unit is known. They make the strong assumption that this information can be given by the interviewed sample units. For agricultural statistics, this assumption implies that each of the selected farmers knows which frames include his farm.

As seen in paragraphs 3, 4 and 5, the second assumption made by Mecatti and Singh (2014), that the union of the collection of frames covers the target population is seldom realistic, unless one of the frames is an area frame, in fact, an area frame is complete by definition. An area frame should be used if another complete frame is not available, an existing list of sampling units changes very rapidly, an existing frame is out of date, an existing frame was obtained from a census with low coverage or a multiple purpose frame is needed for estimating many different variables linked to the territory (agricultural, environmental etc.) (Carfagna and Carfagna, 2010).

A disadvantage of area frames is the sensitivity to outliers and the instability of estimates. The most widespread way to avoid instability of estimates and to improve their precision is adopting a multiple frame sample survey design. For agricultural surveys, a list of very large operators and of operators that produce rare items is combined with the area frame. If this list is short, it is generally easy to construct and update. Some studies are still needed for assessing the difficulty in identifying the farmers through an area frame, when interviews have to be conducted for collecting data concerning socio-economic variables and where respondents live far from the selected area units. The computation of the average time needed for identifying the farmers and the risk of missing data, under the different conditions, request additional research.

11 Concluding remarks

In this paper, the traditional approach for generating a master sampling frame for agricultural statistics and the impact of incomplete or out of date sampling frames were analysed, as well as some approaches for updating the sampling frames. The subject of linking frames for agricultural surveys was explored. 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 makes 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.

An alternative to dealing with a list frame building process is to use 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 in order to avoid bias. Since using multiple frames represents 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.

References

Abaye, A. T. (2013) Master sampling frames for agricultural and rural statistics in Ethiopia, Proceedings of The Sixth International Conference on Agricultural Statistics

Benedetti R., Bee M., Espa R., Piersimoni F., eds. (2010) *Agricultural Survey Methods*. Chichester, UK, Wiley. 434 pp.

Berntsen E., Viviano C. (2011) La progettazione dei censimenti generali 2010-2011: la rilevazione di controllo della copertura e qualità del prototipo di registro statistico delle aziende agricole (Clag) e la riconciliazione con la Base integrata delle fonti amministrative (Bifa), *Istat working papers*, n.1 2011 http://www.istat.it/it/files/2011/06/Istat_Working_Papers_1_2011.pdf

Carfagna, E. and Carfagna, A. (2010) "Alternative sampling frames and administrative data; which is the best data source for agricultural statistics?" In R. Benedetti, M. Bee, R. Espa & F. Piersimoni (eds.) *Agricultural Survey Methods*. Chichester, UK, Wiley. 434 pp.

Carfagna, E. , Pratesi M., Carfagna, A. (2013) Methodological developments for improving the reliability and cost-effectiveness of agricultural statistics in developing countries, the *59th World Statistical Congress, Special Topic Session (STS043)* Using geospatial information in area sampling and estimation for agricultural and environmental surveys, Hong Kong, 25-30 August 2013 http://www.statistics.gov.hk/wsc/STS043-P1-S.pdf

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

Giovacchini A. (2012) Area and point sampling frames for agricultural statistics, presentation at the High Level Stakeholders Meeting on the Global Strategy - From Plan to Action, FAO, December 2012

Keita N., Gennari P. (2013) Building a Master Sampling Frame by Linking the Population and Housing Census with the Agricultural Census, the *59th World Statistical Congress, Special Topic Session (STS063)* "Role of population and housing and agricultural censuses in the national statistical systems", Hong Kong, 25-30 August 2013.

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

Lavallée, P. and Rivest L.P. (2012), Capture-Recapture Sampling and Indirect Sampling, *Journal of Official Statistics*, 28, n.1, pp.1-27.

Mazziotta M. (ed.) (2013) La valutazione della qualità. Atti del 6° Censimento Generale dell'Agricoltura, Istituto nazionale di statistica, Roma, Italy.

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

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

Santos, D., Freitas, M., Lila, M., Arantes, S., Dantas, T., Santos, V. (2013) Brazilian agricultural survey system: a description of sampling methods, *Proceedings of The Sixth International Conference on Agricultural Statistics*.

Wallgren A., Wallgren B. (2010) "Using Administrative Registers for Agricultural Statistics" in Benedetti, Bee, Espa, Piersimoni (eds.), *Agricultural Survey Methods*

Winkler W.E and Thibaudeau Y. (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

Yitzkov, T. and Azaria, H. (2003) Record Linkage in an Integrated Census, *FCSM 2003 Research Conference*, Washington DC.