Master sampling frames for agricultural, rural and agro-environmental statistics, methodological and practical issues

Elisabetta, Carfagna
Department of Statistical Sciences, University of Bologna
Via Belle Arti 41, 40126, Bologna, Italy
elisabetta.carfagna@unibo.it

DOI: 10.1481/icasVII.2016.f36d

ABSTRACT

Methodological and practical problems have to be faced when building a master sampling frame for agricultural, rural and agri-environmental statistics. This paper addresses some of them, focusing on quality and coverage issues and on the impact of increasing computational ability to handle massive data sets on the generation and updating of master sampling frames. Advantages, disadvantages and requirements of the combination of different kinds of frames and the main methods for linking frames at the design stage and at the estimation stage are analysed. A proposal for increasing the efficiency of the allocation of the sample units to the different combined frames is also discussed.

Keywords: Master sampling frame, Multiple frames, Single and two-stage estimators

1. Introduction

In this paper, we present an analysis of methodological and practical problems to be faced when building a master sampling frame for agricultural, rural and agri-environmental statistics. We start from the traditional approach for generating a master sampling frame for agricultural statistics and analyse the effect of incomplete or out of date sampling frames. In section 3, the impact of increasing computational ability to handle massive data sets on the generation and updating of master sampling frames is discussed. Then, other kinds of master sampling frames are taken into consideration (section 4). Section 5, presents a review of the main methods for linking frames at the design stage and at the estimation stage, focusing both on single-stage and two-stage estimators. In

section 6, we talk about the use of area sampling frames for collecting crop and agri-environmental data, the advantages, disadvantages and requirements when list frames are combined with area frames, with single-stage, as well as with two-stage estimators, and the difficulties in the identification of the farms selected through the area frame, according to the kind of area frame and adopted technological tools. Section 7 focuses on a method for improving the efficiency of the allocation of the sample units to the different combined frames. Finally, some conclusions are drawn.

2. The traditional approach for generating a master sampling frame

A master sampling frame is a sampling frame that provides the basis for all data collections through sample surveys and censuses in a certain sector, allowing to select samples for several different surveys or different rounds of the same survey, as opposed to building an ad-hoc sampling frame for each survey. The aims of the development of a master sampling frame are: avoiding duplication of efforts, reducing statistics discrepancies, connecting various aspects of the sector, allowing the analysis of the sampling units from the different viewpoints, and having a better understanding of the sector. 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. Data are collected through mail, email, personal interviews, computer assisted personal interviews, computer assisted telephone interviews, or the web. The census allows generating the list frame that is updated on the basis of administrative data, in the period between two successive censuses and is used for all kinds of sample surveys of farms; thus, it could be considered as a master sampling frame for agricultural statistics. An assessment of the quality of the data collected allows deciding if and how to use this list as a master sampling frame. For example, at the end of the data collection of the Italian agricultural census, a sample survey for assessing the quality of collected data was designed (Mazziotta, 2013). 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. This assessment showed that the complete enumeration census systematically underestimates the main structural variables that 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, reducing the efficiency of the stratification.

3. Impact of increasing computational ability on the generation and updating of master sampling frames

The unbiasedness of this kind of list frame depends on the level of under-coverage and over-coverage of the list at the census date and on the quality of data and the process used for updating the list after the census date. This updating process has become easier, due to great improvements in data base management, including geographic databases (GIS). Moreover, methodological developments for deterministic as well as for probabilistic record linkage have considerably increased the capacity to identify the same record in different lists. For the Italian agricultural census, a very accurate assessment of the coverage was carried out (Mazziotta, 2013) 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). The 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 decreased 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.

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 suggest 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.

4. Other kinds of master sampling frame

Other approaches have been developed for creating master sampling frames. In several countries, the population census is conducted using an administrative structure in which cartographic or other mapping materials are used to divide the country into enumeration areas. The sampling frame is the list of enumeration areas. In agricultural sample censuses and surveys, a sample of enumeration areas is selected, the list of households in selected enumeration areas is created and a sample is extracted from each of these lists, following a two stages sample design. In

many countries, a sample agricultural census is conducted: some enumeration areas are randomly selected and screened for farms. The resulting sampling frame consists of the agricultural census enumeration areas. These approaches present coverage problems at least of the entity of the complete enumeration agricultural census described before. A proposal by FAO and UNFPA aims at avoiding to face the cost of the agricultural census: the list of farms or agricultural households is 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, 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). In some countries, the list of the farms is based on administrative sources, such as business registrations or tax collections. A big disadvantage of the administrative sources is that they may not include the total population, especially units below a threshold required to be registered or pay taxes. In other words, while they will be inclusive of commercial farms, are not likely to include small-scale farms and subsistence farming units (see Carfagna and Carfagna, 2010).

5. Linking frames at the design stage and at the estimation stage

When the coverage and the accuracy of the structural characteristics are not high, alternative approaches can be followed: creating a sampling frame integrating different lists (design level), combining estimates from different lists (estimator level), using an area frame, combining an area frame with one or more list frames. The first option foresees that different lists concerning the same population are used for creating the sampling frame. In such a case, one single frame is created on the basis of two or more lists. In order to get one list combining more than one, records have to be matched. This is not an easy task because farms can appear with different pieces of information in the different lists, and sometimes only partial or wrong information is available. A wide literature has been developed on record linkage, focusing on deterministic and probabilistic rules for matching; moreover, the capacity of storing and managing databases is increased impressively. However, the coverage of the sampling frame is strongly influenced by the quality of the combined lists. Lists with limited coverage or out of date information can create difficulties in the record linkage process, increase the over-coverage and give little contribution to reduce the under-coverage of the sampling frame. Unless the different lists contribute with essential information to complete the frame and the record matching gives extremely reliable results, the frame will be still incomplete and with many duplications (see Carfagna and Ferraz, 2015).

Another option is treating the different lists separately and selecting samples from each list. All observations can be treated as though they had been sampled from a single frame, with modified weights for observations in the intersection of the lists (single-stage estimation). The basic idea is that a multiple frame sample can be viewed as a special case of selecting two or more samples independently from the same frame. As stated by Kalton and Anderson (1986), when a sample is drawn from two or more overlapping frames, the chance of an element being selected depends on the number of frames on which it appears. Compensation for the varying inclusion probabilities of different population elements may be made, by means of a weighting adjustment in the analysis,

such as assigning sample element weights made inversely proportional to their inclusion probabilities. Kalton and Anderson (1986) and Skinner (1991) proposed an unbiased estimator that does not require determining the common units of samples from the different frames. Mecatti (2007) and Mecatti and Singh (2014) also gave a contribution to the development of single-stage estimators proposing their multiplicity estimator. Like the other single-stage estimators developed previously, the Mecatti and Singh estimator has two crucial requirements: the multiplicity of each sample unit is known and the union of the collection of frames covers the target population. Mecatti and Singh (2014) assume that the information on the multiplicity 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. The assumption that the union of the collection of frames covers the target population is seldom realistic, even in developed countries. Indeed, if the aim is providing a rough estimate of main agricultural items, the bias introduced by a limited under-coverage tends to be not particularly high, since generally it concerns mainly small farms, whose contribution to the total of main items is limited. However, the bias can be higher and difficult to remove for minor and special agricultural items. Moreover, small farms are important if we want to have an overview of the trends in rural areas. Another way of taking advantage of various frames at the estimator level is adopting an estimator that combines estimates calculated on non-overlapping sample units belonging to the different frames with estimates calculated on overlapping sample units (two-stage estimation). Two-stage estimators do not require the knowledge of the multiplicity for selected units, but assume that the union of the collection of frames covers the target population. Some two-stage estimators need the identification of identical units only in the overlap samples and some others have been developed for cases in which these units cannot be identified (see Fuller and Burmeister 1972). Both single-stage and two-stage estimators do not require record matching of listing units of the different frames (a process that is notoriously error prone when large lists are used). Generally, complex designs are adopted in the different frames to improve the efficiency and this affects the estimators. Lohr and Rao (2006) proposed optimal estimators and pseudo maximum likelihood estimators when two or more frames are used. Ferraz and Coelho (2007) investigated the estimation of population totals incorporating available auxiliary information from one of the frames at the estimation stage, for the case of a stratified dual frame survey; for a review of multiple frame estimators see Carfagna (2001) and Carfagna and Carfagna (2010).

6. Combining lists and area frames, advantages, disadvantages and requirements

Combining a list and an area frame is a special case of multiple frame sample surveys in which sample units belonging to the lists and not to the area frame do not exist. This approach is very convenient when the list contains units with large (thus probably more variable) values of some variables of interest and the survey cost of units in the list is much lower than in the area frame.

Ground data collection through an area frames is the most reliable way for collecting crop data and some agri-environmental data linked to the land, like the ones included in the field data collection form 2015 of the European land use and cover area frame survey (LUCAS). These data allow computing the following indicators: land cover/land use/change, parcel size, cropping system/land management, irrigation, landscape elements, associated trees and shrubs, soil erosion/soil quality. Ground positioning systems (GPS), aerial images, aerial photos (also photo-interpreted and stored on a PDA, Google Earth, Geographic information Systems (GIS) have considerably modified the data collection process and increased the quality of data.

If economic and rural characteristics and/or agri-environmental indicators related to the farm management are relevant for a country, the ground observation through an area frame is not sufficient and the farmers have to be selected and interviewed. Moreover, when the area frame is combined with one or more list frames, the presence on the lists of the farms selected through the area frame has to be assessed for most estimators.

The main typologies of area frames are segments, with or without physical boundaries, and clustered and un-clustered points. When segments are adopted, the fields totally or partially included in the segments can be used for identifying the corresponding farms; then, from the estimation viewpoint, the traditional open, closed and weighted estimators can be taken into consideration. The number of farms indirectly selected through a segment depends on the number of parts of farms included in the segment; thus, it changes from segment to segment and only an expected number of farms can be prefixed by selecting the segment size. If clustered or un-clustered points are selected, the field corresponding to the point identifies the farm.

The challenging part is collecting the data of the farm corresponding to the field. This task is difficult when the farmers live in villages far from the land. When un-clustered point sampling is adopted, the identification of the farmer is more cumbersome because the next farmer to be identified is far away. Close farmers are easier to identify, since one of them can give some information on the others. Sometimes, point sampling of farms in a segment is carried out, in order to select only a subset of the farms totally or partially included in the segment. This approach is appropriate where the optimum segment size for collecting area and yield information in the fields is larger than the optimum segment size for farmers' interviews. This happens where the farm size is small. Point sampling in the segments also allows prefixing the number of farms selected in each segment, in case point sampling with replacement is adopted (the same farm can be selected by more than one point). This is a big advantage for the sample allocation to the frames.

7. Sample allocation

Under a linear cost function, the optimum share of the total sample to be allocated to each frame can be determined, in order to optimize the precision of the total estimate. However, the optimum sample allocation depends on the variances of domains, which are generally unknown before the survey. An adaptive sequential approach could be adopted for determining the allocation during the survey. Consider that adaptive sequential sample designs are very efficient because the sample selection depends on previously selected units and the stopping rule is based on the estimate. Unfortunately, sequential sample designs are biased, for the same reasons. Thompson and Seber (1996, pages 189-191) faced the problem of sample allocation without previous information on the variability inside strata suggesting a stratified random survey in two phases or, more generally, in k phases. In our case, the strata represent the strata in the different sampling frames. At the k-th phase, a complete stratified random sample is selected, with sample sizes depending on data from previous phases. Then the conventional stratified estimator, based on the data from the k-th phase, is unbiased for the population total Y. The key to design unbiasedness of such an estimator is that each of the estimators is design unbiased and that the weights are fixed in advance and do not depend on observations made during the survey, which implies that, at whatever k-th phase, each of the strata needs to be sampled. These elements guarantee unbiased but not very efficient estimates. Carfagna and Marzialetti (2009), proposed the adoption of an adaptive sequential sample selection with permanent random numbers, which allows optimizing the sample allocation to the different strata and the use of optimum weights for estimating the population total. This procedure foresees that one sample unit is selected at each step, the standard deviations of the

domains are computed and the next sample unit is assigned to the stratum where the sample size is farthest below the size assigned by Neyman's allocation. In the case of the sample allocation to two or more sampling frames, a less cumbersome k-step procedure with permanent random numbers, where k is equal to a small (2 or 3) number of steps is more appropriate. A permanent random number is assigned to all sampling unit in each domain (each stratum of each sampling frame). Then, a first random sample of sampling units is selected. The main aim of this first sample is generating a first estimate of the standard deviations in the domains, which are used for determining the optimum allocation of the second step sample and the optimum weights for combining the estimates from the various lists, then the process can be repeated.

8. Concluding remarks

The quality of the data collected by a complete enumeration census of agriculture should be checked before using the list of farms generated by the census as a master sampling frame, since the under-coverage is about 20% in developed countries. The impressive progress in managing big amount of data and the use of georeferenced data have considerably improved the quality of the updated list; however, this kind of update does not eliminate the under-coverage and can increase the over-coverage, creating several data collection problems. Creating a muster sampling frame integrating different kinds of lists, taking advantage of the improvements in record linkage can be an alternative. However, unless the different lists contribute with essential information to complete the frame and the record matching gives extremely reliable results, the frame will be still incomplete and with many duplications. Another option is treating the different lists separately and selecting samples from each list, using a single-stage or a two-stage estimator. The single-stage estimators have crucial requirements which are seldom satisfied, whiles two-stage estimators facilitate the use of different and complex sample designs in the different lists, increasing the efficiency of the estimators. Ground data collection through an area frame is the most reliable way for collecting crop data and some agri-environmental data linked to the land; however, if all economic characteristics and/or agri-environmental indicators related to the farm management have to be estimated, the ground observation through an area frame is not sufficient and the farmers have to be selected and interviewed. When un-clustered point sampling is adopted, the identification of the farmer is cumbersome because the next farmer to be identified is far away. The optimum sample allocation to different strata of the sampling frames depends on their variances, which are generally unknown before the survey. An adaptive sequential approach for determining the allocation during the survey increases the efficiency of the estimates.

REFERENCES

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

Carfagna E. (2001), Multiple Frame Sample Surveys: Advantages, Disadvantages and Requirements, in International Statistical Institute, Proceedings, Invited papers, International Association of Survey Statisticians (IASS) Topics, Seoul August22-29, 2001, pp. 253-270.

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. and Ferraz C. (2015) Updating sampling frames for agricultural statistics: approaches, challenges and issues", the 60th World Statistical Congress, Proceedings, Specialized Topic Session, Rio De Janeiro, 26-31 July 2015, International Statistical Institute.

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, Proceedings, Special Topic Session, Hong Kong, 25-30 August 2013

Carfagna E. and Marzialetti J. (2009) Sequential Design in Quality Control and Validation of Land Cover Data Bases, Journal of Applied Stochastic Models in Business and Industry, Volume 25, Issue 2, 2009, pp. 195-205, DOI: 10.1002/asmb.742, John Wiley & Sons, Ltd.

Ferraz C., Coelho H.F.C. (2007), Ratio Type Estimators for Stratified Dual Frame Surveys, in Proceedings of the 56 session of the ISI, 2007, Lisbon.

Fuller, W.A., & Burmeister, L.F. (1972). Estimators of samples selected from two overlapping frames, Proceedings of the Social Statistics Sections, American Statistical Association, 245-249.

Kalton G. and Anderson D. W. (1986), Sampling rare populations, Journal of the Royal Statistical Society, Ser. A, 149, pp. 65-82

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, Proceedings, Special Topic Session, 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.

Lohr, S., and Rao, J.N.K. (2006), Multiple frame surveys: Point estimation and inference, Journal of American Statistical Association, 101, 1019-1030.

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.

Skinner C. J. (1991) On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys, Journal of the American Statistical Association, vol. 86, No. 415, Theory and Methods, pp. 779-784.

Thompson S.K., Seber G.A.F. (1996) *Adaptive Sampling*, Wiley, New York.