Combining list frames with different kinds of area frame

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

Andrea Carfagna
Independent consultant, Offida, Italy – andreacarfagna@virgilio.it

Abstract

The problem of producing agricultural statistics when incomplete or out of date lists are available is crucial for many countries, which take the decision to combine different lists, in order to increase the coverage. If the frame is still incomplete, an area frames can be used, in order to guaranty complete coverage and, consequently, unbiased estimates. However, area frames generally include outliers and combining area frames with list frames offers the possibility to reduce the instability of estimates, increasing their precision. Main estimators proposed in the literature for combining area and list frames require the identification of list sample units included in the area frame sample. The difficulties connected to this activity is analysed, with reference to the different kinds of area frame. Another problem in the combination of area and list frames is determining the sample allocation to the different frames, for which a solution is proposed, based on a two-stage adaptive sample selection procedure with permanent random numbers. This proposal optimizes the allocation of sampling units to the frames during the sample design, and guaranties unbiased estimates.

Keywords: list frames, area frames, combination of frames, multiple frames

1. Introduction

At country level, annual agricultural statistics are produced following a variety of approaches, like subjective estimates, use of administrative data, sample surveys, and so on.

A traditional approach is carrying out probability sample surveys of farms selected from a list, generated by a census of agriculture and updated on the basis of administrative data, in the period between two successive censuses.

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 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. However, in developed countries, the coverage of the list generated by the census is around 80%; furthermore, the efficiency of this list frame is influenced by the accuracy of the structural information collected by the census that can be used for stratification purposes, that is not very high (for more details, see Carfagna and Ferraz, 2015).

When the coverage and the accuracy of the structural characteristics are not high, alternative approaches can be followed:

1. Creating a sampling frame integrating different lists (design level)
2. Combining estimates from different lists (estimator level)
    a. Single-stage estimators
    b. Two-stage estimators
3. Using an area frame
4. Combining an area frame with one or more list frames.

In this paper, we discuss these alternative approaches, focusing on the combination of an area frame with list frames, to reduce the instability of estimates and increase their precision. Attention is devoted to the difficulties connected to the identification of list sample units included in the area frame sample for using some estimators proposed in the literature. The problem of determining the sample allocation to the frames, when combining an area frame with list frames, and a solution based on a two-stage adaptive sample selection procedure with permanent random number is proposed, in order to optimize the allocation of sampling units to the frames during the sample design and guaranty unbiased estimates. Finally, the role of point sampling in segments is explained, also in relation to the sample allocation.

2. Combining various lists

The first option mentioned in the introduction 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 undercoverage 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 (option 2 in the introduction) is treating the different lists separately and selecting samples from each list. For simplicity, let us consider the case of two frames (*A* and *B*), both incomplete and with some duplications, which together cover the whole population. The frames A and B generate three (2^2-1) mutually exclusive domains: *a* (units in *A* alone), *b* (units in *B* alone), *ab* (units in both *A* and *B*). *NA* and *NB* are the frames sizes, *Na*, *Nb* and *Nab* are the domains sizes.

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 requirement:

1. The multiplicity of each sample unit is known;
2. The union of the collection of frames covers the target population.

Mecatti and Singh 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 (see Carfagna and Ferraz, 2015).

Indeed, if the aim is providing a rough estimate of main agricultural items, the bias introduced by a limited undercoverage 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 Hartley 1962, 1974 and 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).

3. Advantages and disadvantages of area frames

As said before, both single-stage and two-stage estimators assume that the union of the collection of frames covers the target population. This assumption is seldom realistic, unless one of the frames is an area frame, in fact, an area frame is complete by definition.

In order to avoid biased estimates due to under coverage, 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.). Area frame sample designs also allow objective estimates of characteristics that can be observed on the ground, without interviews. Besides, the materials used for the survey and the information collected help reducing non-sampling errors in interviews and are a good basis for data imputation for non-respondents. Finally, the area sample survey materials have become cheaper and more accurate (Carfagna and Carfagna, 2010).

Area frame sample designs also have some disadvantages, such as the cost of implementing the survey program, the necessity of many cartographic materials, the sensitivity to outliers and the instability of estimates. If the survey is conducted through interviews and respondents live far from the selected area unit, their identification may be difficult and expensive, and missing data tend to be relevant.

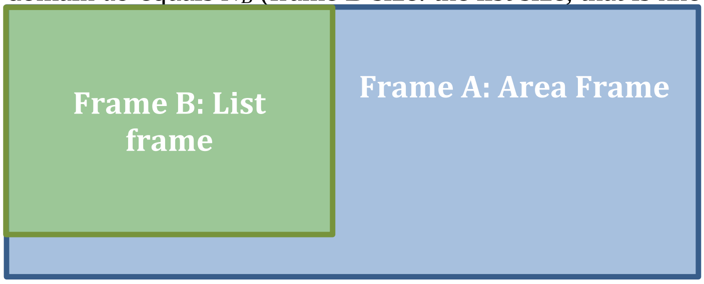
4. Combining list and area frames

The most widespread way to avoid the instability of estimates and to improve their precision is adopting a multiple frame sample survey design. For surveys on economic activities, 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. A crucial aspect of this approach is the identification of the area sample units included in the list frame. When units in the area frame sample and in the list frame are not detected, the estimators of the population totals have an upwards bias.

Sometimes, a large and reliable list is available. In such cases, the final estimates are essentially based on the list sample. The role of the area frame component in the multiple frame approach is essentially solving the problems connected with incompleteness of the list and estimating the incompleteness of the list itself. In these cases, updating the list and record matching for detecting overlapping sample units in the two frames are difficult and expensive operations that can produce relevant non-sampling errors (for more details, see Vogel 1975 and Kott and Vogel 1995).

5. Multiple frame sample surveys with known domain sizes

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 (domain *b* is empty) and the size of domain *ab* equals N_B (frame B size: the list size, that is known).



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. In fact, when *A* is an area frame, we have:

$$\hat{Y} = \hat{Y}_a + p\hat{Y}_{ab}^{A} + q\hat{Y}_{ab}^{B}$$

with variance:

$$Var(\hat{Y}) = Var(\hat{Y}_a) + p^2 Var(\hat{Y}_{ab}^{A}) + (1-p)^2 Var(\hat{Y}_{ab}^{B}) + 2pCov(\hat{Y}_a, \hat{Y}_{ab}^{A})$$

and the value of *p* that minimises the variance is:

$$p_{opt} = \frac{Var(\hat{Y}_{ab}^{B}) - Cov(\hat{Y}_a, \hat{Y}_{ab}^{A})}{Var(\hat{Y}_{ab}^{A}) + Var(\hat{Y}_{ab}^{B})} \quad .$$

The optimum value of *p* depends on the item and can assume very different values for the different variables.

If the precision of the estimate of the total for the overlapping sampling units is low, ($Var(\hat{Y}_{ab}^{A})$ is high) its contribution to the final estimate is low; thus, in some applications, the value of *p* is chosen equal to zero and the resulting estimator is called screening estimator, since it requires the screening and elimination from the area frame of all the area sampling units included in the list sample.

6. 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 *A* and *B*, which are 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 domains. 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.

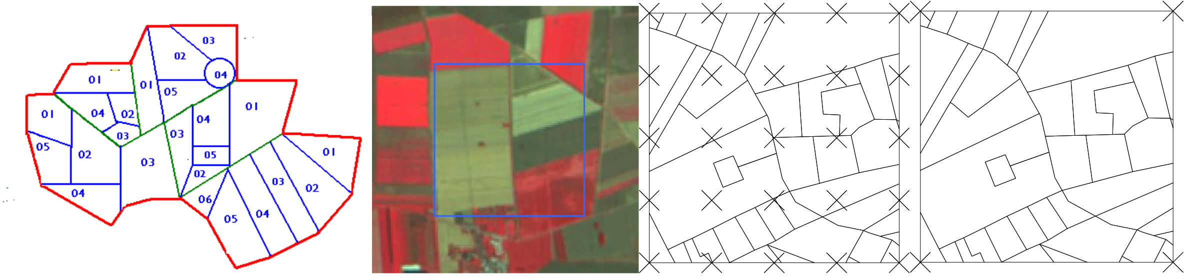
Carfagna and Marzialetti (2009), proposed the adoption of an adaptive sequential sample selection with permanent random numbers (Ohlsson‘95), which allows optimizing the sample allocation to the different strata and the use of optimum weights for estimating the population total. This procedure foresees 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 an area and one or more list frames, we propose adopting a less cumbersome *k*-step procedure with permanent random numbers, where k is equal to a small (3 or 4) number of steps. A permanent random number is assigned to all sampling unit in each domain. 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 value of *p* and the optimum allocation of the second step sample, and the process repeats until the total budget is reached.

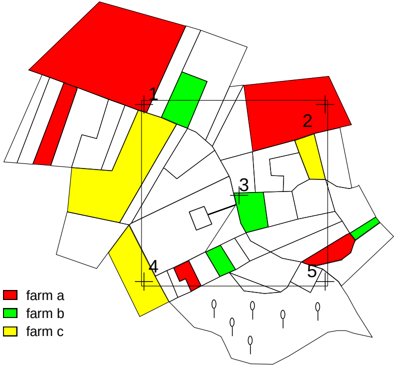
Further research topics concern the identification of the optimum number of steps to be used and the number of sample units to be sampled at each step.

8. Various kinds of area frame

Various kind of area frames can be adopted. Main typologies are segments, with or without physical boundaries, and clustered and un-clustered points, in the respective order in the figure below.



From the point of view of the combination of area and list frames, the crucial point is the difficulty in identifying the farms in the area sample and assessing their presence on the list.
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, in order to assess if the farm is included in the list (or lists). This task is difficult when the farmers live in villages far from the land. When un-clustered point sampling is adopted, the probability of identifying the farmer is lower 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, as in the following figure:



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.

9. Concluding remarks

In this paper, we have addressed the problem of producing agricultural statistics when incomplete or out of date lists are available, focusing on the combination of different lists with single-stage and two-stage estimators. The combination of an area frame with list frames guarantees complete coverage and, thus, unbiased estimates; moreover, it simplifies both kinds of estimators. In fact, with an area frame, the multiplicity of farms is at least one (the area frame) in the single-stage estimator and the size of the domains in the two-stage estimator is known. The kind of area frame affects the difficulty in the identification of the farms selected through the area frame, which is more cumbersome when un-clustered point sampling is adopted.

The optimal allocation of sample units in the area and list frames depends on the variability in the frames; that is unknown when the sample is designed. Thus, we have proposed a *k*-step adaptive procedure with permanent random numbers, which allows optimizing the sample design while carrying out the survey, and guarantees unbiased estimates.

Point sampling with replacement in selected segments facilitates the sample allocation because allows prefixing the number of selected farms in selected segments.

References

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, Specialized Topic Session (STS073)* "Agriculture Master Frame: beyond creation", organized by Flavio Bolliger and Andrea Diniz da Silva, Rio De Janeiro, 26-31 July 2015. International Statistical Institute.

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.

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

Hartley, H.O. (1974), Multiple frame methodology and selected applications, *Sankhya*, C, 36, 99 -118.

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

Kott P. S., Vogel F. A. (1995), Multiple-frame business surveys, in Cox, Binder, Chinnapa, Christianson, Colledge, Kott (Eds.), *Business survey methods*, Wiley, New York, pp. 185-201.

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

Mecatti, F. (2007) A single frame multiplicity estimator for multiple frame surveys, *Survey Methodology*, 33, 151-58

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.

Ohlsson E. (1995) Coordination of Samples Using Permanent Random Numbers, in *Business survey methods*, Cox B. G., Binder D.A., Chinnapa B.N., Christianson A., Colledge M.J., Kott P.S. (Eds.), Wiley, New York, 153-169.

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 an Methods, pp. 779-784.

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

Vogel F. A. (1975) Surveys with Overlapping Frames - Problems in Applications, *Proceeding of the Social Statistics Section, American Statistical Association*, pp. 694-699.