Multiple Frame Sample Surveys: Advantages, disadvantages and Requirements

Elisabetta Carfagna
Department of Statistics, University of Bologna
Via Belle Arti 41 – 40126
Bologna, Italy
carfagna@stat.unibo.it

1. Introduction

One of the most important practical problems in conducting sample surveys is that lists that can be used for selecting the samples are generally incomplete or out of date. Therefore, sample surveys can produce seriously biased estimates of the population parameters. Updating a list is a difficult and very expensive operation that has partially become easier due to the recent advances in managing databases. In any case, the single most important and expensive factor to be considered for updating a list is the data collection effort.

Different lists concerning the same population are sometimes used for obtaining the list of sampling units, called sampling frame. It is assumed that the union of the different frames cover the whole population. For example, a list obtained from a census carried out some years before the sample survey could be updated and integrated by using administrative data. In such a case, one single frame is created, although on the basis of two or more lists. This approach should be undertaken only if the different lists contribute with essential information to complete the frame and the record matching gives extremely reliable results; otherwise, the frame will be still incomplete and with many duplications. Another option is using the different lists in a multiple frame approach, that 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.

2. The multiple frame approach

Some relevant examples of the combined use of different frames can be found since 1949 (the sample survey of retail stores conducted by the US Bureau of the Census). Later, in 1962, Hartley developed the basic theory of multiple frame sampling. Hartley considered dividing the population into mutually exclusive domains defined by the sampling frames and their intersections, and proposed a methodology that allows utilizing any number of frames. Two important assumptions have to be made: i) Completeness: every unit in the population of interest should belong to at least one of the frames; ii) Identifiability: it should be possible to record, for each sampled unit, whether or not it belongs to one or more of the other frames.

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*). N_A and N_B are the frames sizes, N_a, N_b and N_{ab} are the domains sizes. Generally, the three domains cannot be sampled directly, since samples of sizes n_A and n_B have to be selected from frames *A* and *B*. Thus n_a, n_{ab}^A, n_{ab}^B and n_b (the subsamples of n_A and n_B respectively which fall into the domains *a*, *ab* and *b*) are random numbers and a post-stratified estimator has to be adopted for the population total. For simple random sampling in the two frames, in case all the domain sizes are known, a post-stratified estimator of the population total is the following:

$$\hat{Y} = N_a \bar{y}_a + N_{ab}(p\bar{y}_{ab}^A + q\bar{y}_{ab}^B) + N_b \bar{y}_b, \tag{1}$$

where *p* and *q* are non-negative numbers with $p + q = 1$; $\bar{y}_a$ and $\bar{y}_b$ denote the respective sample means of domains *a* and *b*; finally, $\bar{y}_{ab}^A$ and $\bar{y}_{ab}^B$ are the sample means of domain *ab*, relative, respectively, to subsamples n_{ab}^A and n_{ab}^B. The means $\bar{y}_a$ and $\bar{y}_{ab}^A$ are replaced by $\bar{y}_A$ (the sample mean relative to the whole n_A sample) if either n_a=*0* or n_{ab}^A=*0*; likewise $\bar{y}_b$ and $\bar{y}_{ab}^B$ are replaced by

$\bar{y}_B$ if either $n_b=0$ or $n_{ab}^B=0$. $N_a\bar{y}_a$ is an estimate of the incompleteness of the list. Hartley (1962) proposed to use the variance for proportional allocation in stratified sampling as approximation of the variance of the post-stratified estimator of the population total $\hat{Y}$ (ignoring finite population corrections):

$$Var(\hat{Y}) \approx \frac{N_A^2}{n_A}\left[\sigma_a^2(1-\alpha)+p^2\sigma_{ab}^2\alpha\right]+\frac{N_B^2}{n_B}\left[\sigma_b^2(1-\beta)+q^2\sigma_{ab}^2\beta\right], \qquad (2)$$

where s_a^2 s_b^2 and s_{ab}^2 are the population variances within the three domains, moreover $\alpha = N_{ab}/N_A$ and $\beta = N_{ab}/N_B$. It is well known that this approximation is an underestimation of the variance of the post-stratified estimator, since it ignores the increase in variance that arises because the sample units do not distribute themselves proportionally in the different domains. Underestimation is small only if the average sample size per domain is sufficiently large and errors in the domain sizes can be ignored. Under a linear cost function, the values for p, n_A/N_A and n_B/N_B minimising the estimator variance can be determined (see Hartley, 1962).

The knowledge of the domain sizes is a very restrictive assumption that is seldom verified. Often, domain sizes are only approximately known, due to the use of out of date information and lists, that makes difficult to determine whether a unit belongs to any other frame. In such a case the estimator of the population total given in equation (1) is biased and the bias remains constant as the sample size increases. The estimator of the variance given in equation (2) underestimates the true error of $\hat{Y}$ (since it doesn' t contain the contribution of the bias to the error) and the mean square error should be computed. Various authors, such as Hartley (1962 and 1974) and Fuller and Burmeister (1972), proposed some estimators of the population total when the domain sizes are not known.

We can conclude that a multiple frame approach should be adopted only if the different frames contribute with essential information. Moreover, the number of used frames should not be high, otherwise the sample size per domain would be small, the domain sizes would probably be only approximately known and the population total estimator could be seriously biased. Finally, with many frames, some of which out of date, record matching is very difficult and errors in record matching are another source of bias.

3. Combining a list and an area frame

An area frame (a set of geographical areas) is always complete, and remains useful a long time. The completeness of area frames suggests their use in many cases, for example: a) if other complete frame is not available; b) if an existing list of sampling units change very rapidly; c) if an existing frame is out of date; d) if an existing frame was obtained from a census with a low coverage; e) if a multiple purpose frame is needed for estimating many different variables (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 to reduce non sampling errors in interviews and are a good basis for data imputation for non-respondents; finally, the area sample survey materials are becoming cheaper and more accurate.

Area frame sample designs also have some disadvantages, such as the cost of implementing the survey program, the necessity of more 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. The most widespread way to avoid the instability of estimates and to improve their precision is adopting a multiple frame sample survey. 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. When units in the area frame sample and in the list 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 of the multiple frames 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 (for more details, see Kott and Vogel 1995).

Combining a list and an area frame is a special case of multiple frame sample surveys with known domain sizes; in fact, 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, that is known). Thus the total of domain b equals zero and the estimator of the population total in equation (1) becomes:

$$\hat{Y} = N_a \bar{y}_a + N_{ab}(p\bar{y}_{ab}^A + q\bar{y}_{ab}^B) \,. \qquad (3)$$

Since $N_B = N_{ab}$, $\boldsymbol{b} = 1$ and $\sigma_B^2 = \boldsymbol{s}_{ab}^2$, equation (2) becomes:

$$Var(\hat{Y}) \approx \frac{N_A^2}{n_A}\left[\sigma_a^2(1-\alpha) + p^2\sigma_{ab}^2\alpha\right] + \left[\frac{N_{ab}^2}{n_{ab}^B} q^2\sigma_{ab}^2\right]. \qquad (4)$$

Hartley computed the variance of the population total for the optimum design, that is using the values for p, n_A / N_A and n_B / N_B which minimize the estimator variance under a linear cost function, and made a comparison with the variance of a post-stratified estimator computed from a simple random sample of size $n_A^* = C / c_A$ selected from the area frame only (called weighted estimator). Then different values were considered for the following parameters: σ_{ab}^2/σ_a^2, c_B/c_A and N_B/N and noticed that the variance reduction with the optimum design is high when the ratio σ_{ab}^2 $/\sigma_a^2$ is high and the ratio c_B/c_A is low. So, it is very convenient to combine a list and an area frame in a multiple frame approach when the list contains units with large (thus probably more variable) units and the survey cost of units in the list is much lower than in the area frame. However, we have to point out that only variable costs have been taken into account and fixed costs tend to be higher in the multiple frame approach due to the more complex sample design and the record matching procedure.

From a general viewpoint, whatever the sample design in the two frames, using the Horvitz-Thompson estimators of the totals of the different domains, the estimator of the population total is given by:

$$\hat{Y} = \hat{Y}_a + p\hat{Y}_{ab}^A + q\hat{Y}_{ab}^B + \hat{Y}_b \,. \qquad (5)$$

Since sample selection is independent in the two frames, the following covariances are zero:

$$Cov(\hat{Y}_a, \hat{Y}_b), Cov(\hat{Y}_a, \hat{Y}_{ab}^B); Cov(\hat{Y}_b, \hat{Y}_{ab}^A); Cov(\hat{Y}_{ab}^A, \hat{Y}_{ab}^B)\,, \qquad (6)$$

the variance of population total in equation (5) is:

$$\begin{aligned} Var(\hat{Y}) &= Var(\hat{Y}_a) + p^2 Var(\hat{Y}_{ab}^A) + (1-p)^2 Var(\hat{Y}_{ab}^B) + Var(\hat{Y}_b) + \\ &+ 2pCov(\hat{Y}_a, \hat{Y}_{ab}^A) + 2(1-p)Cov(\hat{Y}_b, \hat{Y}_{ab}^B). \end{aligned} \qquad (7)$$

Thus, the value of p that minimizes the variance in equation (7) is:

$$p = \frac{Var(\hat{Y}_{ab}^B) + Cov(\hat{Y}_b, \hat{Y}_{ab}^B) - Cov(\hat{Y}_a, \hat{Y}_{ab}^A)}{Var(\hat{Y}_{ab}^A) + Var(\hat{Y}_{ab}^B)} \,. \qquad (8)$$

The optimum value for p is directly related to the precision of $\hat{Y}_{ab}^A$. When list A is an area frame, $\hat{Y}_b$ in equation (5) is zero as well as $Cov(\hat{Y}_b, \hat{Y}_{ab}^B)$ in equation (8). In most applications, the value of p is chosen equal to zero in equation (5) and the resulting estimator is called screening estimator, since it requires the screening of all the area sampling units included in the list and its variance reduces to: $Var(\hat{Y}) = Var(\hat{Y}_a) + Var(\hat{Y}_{ab}^B)$. Armstrong (1979) made a comparison between the weighted, screening and Hartley estimators and noticed that the weighted estimator generally gives lower estimates. Moreover, the screening estimator is much more efficient than the weighted estimator, while the Hartley estimator is slightly more efficient than the screening estimator.

4. Multiple Frame Sampling for Multivariate Stratification

Surveys are often designed for estimating means and totals of many variables and several stratifying variables are available. The usual approach is to use some multivariate stratification scheme that represents a compromise solution for the different purposes. An alternative approach is selecting independent subsamples, each one selected by stratified sampling with respect to just one of the stratifying variables, and then combining, in the estimation process, data from separate subsamples using multiple frame techniques. Thus, in this approach, the use of multiple frame techniques aims at an efficient use of different auxiliary variables and not at the use of different incomplete frames of the same population, or at the association of a complete but inefficient frame with an incomplete but efficient one. A simulation study undertaken allowed a comparison of some usual techniques for stratification with the multivariate stratification based on multiple frames. The result was that significant gains in efficiency can arise when stratifying variables are highly correlated with one or more survey variables and their mutual intercorrelation are fairly low (Skinner et al., 1994).

REFERENCES

Armstrong B. (1979), Test of multiple frame sampling techniques for agricultural surveys: New Brunswick, 1978, *Proc. of the Section on Survey Research Methods, A. S. A.*, pp. 295-300.

Fuller W.A., Burmeister L.F. (1972) Estimators for samples from two overlapping frames, *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 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*, vol.36, series C, Pt.3, pp.99-118.

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.

Skinner C.J., Holmes D.J., Holt D.(1994) Multiple Frame Sampling for Multivariate Stratification, *International Statistical Review*, 62, 3, pp.333-347.

RÉSUMÉ

On analyse l’utlisation conjointe de plusieurs bases de sondage pour constituer une base multiple qui puisse fournir une solution à l’incomplétitude des bases de sondage sur liste. On donne un aperçu des avantages et des limitations des bases de sondage multiples, ainsi que des problèmes liés aux l’estimateurs. Une attention particulière est dédiée au cas où une des bases de sondage est une collection d’unités géographiques. L’utilisation de bases de sondage multiples pour des enquêtes à plusieus objectifs est également commentée.