Multiple Frame Sample Surveys: Advantages, disadvantages and Requirements (Part of the presentation)

Elisabetta Carfagna

University of Bologna

Department of Statistics

International Statistical Institute, Invited papers,

Seoul August 22-29, 2001

One of the most important practical problems of sample surveys:
lists are often incomplete or out of date

- Consequence: sample surveys can produce seriously biased estimates of the population parameters
- Updating a list is a difficult and very expensive operation
- It has partially become easier due to the recent advances in managing databases
- The single most important and expensive factor to be considered for updating a list is the data collection effort

One way for obtaining a complete and updated sampling frame (the list of sampling units)

- Using different lists concerning the same population
- It is assumed that the union of the different frames covers the whole population
- One single sampling frame is created on the basis of two or more lists
- 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
- 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Another way to overcome the problem of incomplete and out of date lists is using the different lists in a multiple frame approach:

- Adopting an estimator that combines estimates calculated on non-overlapping sample units belonging to the different frames with estimates calculated on overlapping sample units
- 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

Multiple frame approach

- Hartley considered dividing the population into mutually exclusive domains defined by the sampling frame 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.

Multiple frame approach

- 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.

Estimator of 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:

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

- 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*;
- $\bar{y}_{ab}^A$ 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 *na=0* or n_{ab}^A *=0*;
- likewise $\bar{y}_b$ and $\bar{y}_{ab}^B$ are replaced by $\bar{y}_B$ if either *nb=0* or n_{ab}^B *=0*.
- $N_a \bar{y}_a$ is an estimate of the incompleteness of the list.

Variance of the population total estimator

- 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 (ignoring finite population corrections):

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

- where σ_a^2 σ_b^2 and σ_{ab}^2 are the population variances within the three domains, moreover $\alpha = N_{ab}/N_A$ and $\beta = N_{ab}/N_B$.
- Under a linear cost function, the values for p, n_A / N_A and n_B / N_B minimizing the estimator variance can be determined (see Hartley, 1962).

Problems of multiple frames

- 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 post-stratified estimator of the population total is biased and the bias remains constant as the sample size increases;
- The variance of the post-stratified estimator of the population total underestimates the true error (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.

Other problems of multiple frames

- A multiple frame approach should be adopted only if the different frames contribute with essential information.
- The number of used frames should not be high, otherwise:

1. The sample size per domain would be small;
2. The domain sizes would probably be only approximately known;
3. The population total estimator could be seriously biased;
4. With many frames, some of which out of date, record matching is very difficult and errors in record matching are another source of bias.

List frames versus area frames

- List frames are very sensitive to obsolescence but very efficient when they are complete, without duplications and updated.
- An area frame is a probability sample survey in which, at least for one sampling stage, the sampling units are land areas (segments, small areas selected by points, line transepts etc.) (General meaning - FAO 1998).
- An area frame is always complete, in whatever year, and remains useful a long time. The completeness of area frames suggests their use in many cases, e.g.:

1. If other complete frame is not available;
2. If an existing list of sampling units change very rapidly;
3. If an existing frame is out of date;
4. If an existing frame was obtained from a census with a low coverage;
5. If a multiple purpose frame is needed for estimating many different variables (agricultural, environmental etc.).

Advantages and disadvantages of area frames

- Advantages of area sample designs;

1. Allow objective estimates of characteristics that can be observed on the ground, without interviews;
2. 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;
3. The area sample survey materials are becoming cheaper and more accurate.

- Disadvantages of area sample designs:

1. The cost of implementing the survey program;
2. The necessity of cartographic materials:
3. The sensitivity to outliers;
4. The instability of estimates;
5. 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.

A special case of multiple frame sample surveys: combining a list and an area frame

- 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 updated and sampled:
- If this list is short, it is generally easy to construct and update;
- Area and list survey estimates are combined to produce the final estimate;
- 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.

Combining a list and an area frame: estimators

- 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 post-stratified estimator of the population total

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

- becomes:

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

- and its variance

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

- since $N_B = N_{ab}$, $\beta = 1$ and $\sigma_B^2 = \sigma_{ab}^2$, becomes:

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

Combining a list and an area frame: efficiency of the optimum sample design

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;

Then he 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):

He considered different values for the following parameters: $\sigma_{ab}^2 / \sigma_a^2$, c_B/c_A and N_B/N and noticed that the variance reduction with the optimum design is high when the ratio $\sigma_{ab}^2 / \sigma_a^2$ 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 large (thus probably more variable) units and the survey cost of units in the list is much lower than in the area frame.

However, 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.