SURVEYS WITH OVERLAPPING FRAMES - PROBLEMS IN APPLICATION

Frederic A. Vogel, Statistical Reporting
Service, U. S. Dept. of Agriculture

Presented at the 135th Annual Meeting of the American

Statistical Association, Atlanta, Georgia, August 25-28, 1975

SURVEYS WITH OVERLAPPING FRAMES - PROBLEMS IN APPLICATION

Frederic A. Vogel, Statistical Reporting
Service, U. S. Dept. of Agriculture

Introduction

A multiple frame survey may be defined as one relying upon the joint use of two or more sampling frames. One of the first uses of two-frame estimation was the U. S. Census Bureau's "Sample Survey of Retail Stores" conducted in 1949 and described by Hansen, Hurwitz and Madow 1/. The theoretical development of the design and estimation in multiple frame surveys was presented by Hartley 2/ in 1962. Since then, others have studied design and estimation problems associated with multiple frame surveys. A paper by Fuller and Burmeister 3/ provides an excellent reference for most of the theoretical work done to date. Most of the work attempts to improve the estimator presented by Hartley.

Multiple frame surveys are subject to all operational problems that plague single frame surveys. However, by their very design, problems unique to multiple frame surveys also occur. These problems arise from the basic assumptions involved in a multiple frame sample design:

a. Every element of the survey population must be included in at least one of the frames.
b. It must be possible to determine for every selected sample unit whether or not it belongs to any other sample frame i.e., the overlap between frames must be determined.

The latter assumption leads to one of the most critical aspects of a multiple frame survey. Sometime during the survey process it is necessary to determine for every sampled unit whether it could have been selected from another frame also being used. The available theory does not tell us how this determination is to be made - it only gives us alternative estimators to use once the determination is made.

The purpose of this paper is to examine problems involved in the overlap determination, and how they can be considered in the estimation process. More specifically, the problems involve those encountered when using the multiple frame concepts in surveys of farm operators.

Concepts

The remainder of the paper will consider problems occurring when the following sample frames are used:

1. An area frame - This is the complete frame or the 100 percent frame. Every farm operator via a sampling unit (segment of land) has a chance to be selected from this frame. This frame is usually the more expensive to use for obtaining survey data.
2. A list frame - This is usually defined to be a list of potential units (names of farm operators) for the population of interest. It may also contain information by which to stratify. In many survey situations, alternative data collection methods may be used which lead to the use of cheaper mail and telephone data collection procedures. However, this frame is usually incomplete and will not provide information for the entire population of interest.

Therefore, the use of multiple frame sampling is applicable for this situation. It allows one to maximize the use of the cheaper, more efficient list frame, yet when in combination with the complete area frame provides efficient estimates for the population of interest.

Two terms to be used are now defined. The area frame sample (the 100 percent frame) must be divided into two domains for multiple frame estimation:

a. Nonoverlap Domain - This domain consists of population units or farms found via the area frame sample that are not in the list frame.
b. Overlap Domain - This domain contains sample units that are also in the list frame. These farm operations in the area frame sample also had a chance to be selected from the list frame.

An unbiased estimator for the population of interest using the area frame is: ${}_{a}\hat{X} = \sum_{h} \frac{{}_{a}N_h}{{}_{a}n_h} \; {}_{a}x'_h$

where $\frac{{}_{a}N_h}{{}_{a}n_h}$ is the reciprocal of the probability of selecting a sample unit in the area frame and x'_h is the sample total for a particular stratum. The area frame estimator can also be written as:

$$ {}_{a}\hat{X} = \sum_{h} \frac{{}_{a}N_h}{{}_{a}n_h} \left({}_{a1}x'_h + {}_{a2}x'_h\right) = {}_{a1}\hat{X} + {}_{a2}\hat{X} \; . $$

Here ${}_{a1}\hat{X}$ is an estimate of the incompleteness of the list frame or the nonoverlap domain of the area frame. Then ${}_{a2}X$ is the area frame estimate of the population also represented by the list frame (overlap domain).

A multiple frame estimator is:
$\hat{X} = {}_{a1}\hat{X} + P \; {}_{a2}\hat{X} + Q_{b}\hat{X}$ where ${}_{b}\hat{X}$ is an estimate of the overlap domain based on the list frame sample and the weights P and Q are such that $P + Q = 1$.

A simplier multiple frame estimator is one where $P = 0$ and $Q = 1$. Then, no information from the area overlap domain is utilized. However, in either case, it is necessary to divide the area frame into the two domains.

Many agricultural surveys are based on multiple frame sample designs. The list frame consists of names of potential farm operators. While the sampling unit is a name, the reporting unit is all land operated by the particular name. The area frame sample units are small areas of land called segments. Each sample segment is screened for farm operations. A sample segment may contain portions of 3-5 farming operations. The names of the farm operators associated with each parcel of land or operation found inside the segment are obtained during the survey.

If costs were of no object, one could obtain a map that outlined the land area associated with every name on the list. If this were overlaid onto the area frame, only land areas not covered by the list would be in the nonoverlap domain.

In practice, it must be assumed that an area of land can be represented by a name. Then, in the multiple frame context, the overlap of land areas represented by both sample frames is identified by matching names associated with the land.

This is probably the most difficult factor involved in a multiple frame survey. Frrors in this determination are not considered in the estimation phase - thus they fall into the area of nonsampling errors. The name matching operation can be completed manually or by a computer method of record matching as described by Fellegi and Sunter 4/. Whichever procedure is used requires certain decision logic about what is a match and what is a nonmatch. Next, some problems that are encountered, different alternatives for defining the domains, and the consideration of the problems in the estimators will be discussed.

Problems

There are two factors contributing to the problems with domain determination or determining whether a farm operation found in the area frame is also in the list frame.

a. One relates to the matter of duplication in the list. It is very difficult to remove duplication from a list frame. Several procedures have been devised for using computers to remove the duplication; however, the problem will continue to exist. The survey procedures for identifying and adjusting survey data when duplication exists in the list frame must be considered in a multiple frame survey, not only for the estimation but also for domain determination.

b. Some larger farming operations, such as partnerships or corporations contain several individuals that may report for the entire operation. These individuals can appear on the list frame either singularly or in combination with other names. This poses a problem in estimation for the list frame. It also poses a problem in determining whether a given operation is overlap with the list frame or not.

The following table illustrates some of the problems encountered when identifying the overlap between the two frames by matching names.

Table 1--Examples that occur when determining the overlap between two sample frames

Problem number	Name(s) associated with land in area frame sample segment	Name(s) in list frame that may represent land in area frame segment
1	Bill, Bob, Joe, and Sam Jones	Sam Jones Bob Jones
2	Bill, Bob, Joe, and Sam Jones	Sam Jones Bob Jones Robert Jones
3	James Smith Bill Smith Milton Brown	Smith and Brown

Problem 1

There are four names associated with a parcel of land in an area frame sample segment. Two of the four names appear in the list frame. Does the parcel of land in the area frame overlap with land operated by Sam Jones, or with Bob Jones, or with both? Can the land be reported twice from the list frame? Not only do we have the problem of determining overlap between the two frames - there is also the possibility of duplication in the list.

In an operational survey, rules must be established so that such problems as above are handled consistently. Three alternate procedures are compared below.

Procedure A

This rule is based on the following assumptions:

a. Each partner will report for the entire operation and correctly identify all of his partners if he is selected from the list.

b. If more than one partner appears somewhere in the list frame, he will be identified.

Since we assume that each partner will report for the entire operation, the parcel of land found in the area frame overlaps the operation represented by the two names on the list. However, there remains the problem of duplication within the list.

Different procedures are available for handling this duplication in the estimation. One is presented by Gurney and Gonzalez 5/ where the number of times a given operation is duplicated is not known. Another method has been developed by Rao 6/ for the case where the number of times an operation can be selected from the frame is known.

It will be assumed we can determine the number of times every selected unit could have been sampled. This is done by matching each name in the list sample with the remaining names in the list frame. Controls are also built into the survey questionnaire to aid in the detection of possible duplication. For example, each respondent is asked whether he is known by any other name or if any other names are associated with his operation.

Rao's procedure was developed for the case where there is no stratification in the frame. His estimator for the list frame would be:

${}_b\hat{X} = \sum^{n} \frac{{}_bN}{{}_bn} \frac{{}_bx_i}{{}_bA_i}$ where ${}_bN$ and ${}_bn$ are the total number of names and number of selected names respectively from the list frame. ${}_bA_i$ is the total number of times a given unit (farm operation) can be selected from the frame. In this example ${}_bA_i = 2$. This estimator is unbiased because we can write:

$${}_b\hat{x} = \sum^{N} \frac{{}_bN}{{}_bn} \frac{{}_bx_i}{{}_bA_i} \cdot t_i \quad \text{where}$$

$t_i = 0$ if the i^{th} name is not selected
$= 1$ if the i^{th} name is selected, and
$E(t_i) = {}_bn/{}_bN$. Then $E({}_b\hat{X}) = \sum^{N} \frac{{}_bx_i}{{}_bA_i}$.

This becomes unbiased if data for the duplicated name is included in the tabulation every time it is selected. If the value ${}_bx_i/{}_bA_i$ is used every time the duplicated unit is selected, the expected value reduces to:

$\sum^{M} {}_bA_i \frac{{}_bX_i}{{}_bA_i} = {}_bX$ where M is the number of unique units in the frame.

The procedure outlined by Rao can be extended to the case where the duplicated names in the list frame are in different strata.

Again, we wish to estimate the population total (X) for the list frame from a sample. The population value can be obtained by summing over the population as follows:

${}_bX = \sum_h \sum_i^{{}_bN_h} \frac{{}_bX_{hi}}{{}_bA_{hi}}$. Here, ${}_bA_{hi}$ is the total number of times a ${}_bX_{hi}$ unit can be selected from the list frame. It is assumed the ${}_bA_{hi}$ factor can be determined correctly from the sample. The duplicated operation is included in the tabulations every time it is selected.

This is simply a rule for assigning a portion of the duplicated operation to each stratum from which it could have been selected. The portion is determined by the weighting factor which is the number of times it could have been selected in a given stratum divided by the total number of times it could have been selected from the list.

The estimator $\sum_h \sum_i^{{}_aN_h} \frac{{}_bN_h}{{}_bn_h} \frac{{}_by_{hi}}{{}_bA_{hi}}$ for the case where the list duplication occurs in different strata can be shown to be unbiased by writing

$\sum_h \sum_i^{{}_bN_h} \frac{{}_bN_h}{{}_bn_h} \frac{{}_by_{hi}}{{}_bA_{hi}}$. Then $E(\hat{X}) = X$ because the sample is selected independently within strata and each duplicated operation is given the value $\frac{y_{hi}}{A_{hi}}$ no matter how many times it is selected.

The multiple frame estimator is then obtained by adding the list estimator ${}_b\hat{X}$ to the area frame portion, i.e.
$X = {}_{a1}\hat{X} + P \quad {}_{a2}\hat{X} + Q\ {}_b\hat{x}$. The success of this estimator depends on the ability to correctly define the domains in the area frame. If the assumption that each individual will report for the entire partnership does not hold in practice, the estimator becomes biased. This occurs because ${}_{a2}x$ will be estimating for an operation that is not represented by the list. It also means that the ${}_bA_{hi}$ weights are incorrect.

A second rule is used which will minimize the effects of an out-of-date list.

Procedure B

This rule relies on the same assumptions used in Rule A, i.e., each individual partner will report for the entire operation and will correctly identify all his partners. Operational procedures differ however and are illustrated:

a. The total number of partners associated with the parcel of land in the area frame sample unit are identified. The number is designated by ${}_aA_{hi}$.
b. The number of partners associated with the area frame sample unit and that are also on the list frame is determined. This number is ${}_bA_{hi}$ as defined for Procedure A.
c. A weighting factor is determined for assigning a portion of the operation to the list frame and a portion to the area frame. The factor to be applied to the area frame is $1 - \frac{{}_bA_{hi}}{{}_aA_{hi}} = 1 - \frac{2}{4}$.

The factor applied to the duplicated list frame sample units is then $1/{}_aA_{hi} = \frac{1}{4}$
The multiple frame estimator then becomes

$$\hat{X} = \sum_h \sum_i^{{}_an_h} \frac{{}_aN_h}{{}_an_h} \left(1 - \frac{{}_bA_{hi}}{{}_aA_{hi}} \right) {}_ax_{hi} + \sum_h \sum_i^{{}_bn_h} \frac{{}_bN_h}{{}_bn_h} \frac{{}_ax_{hi}}{{}_aA_{hi}}.$$

Note that if ${}_bA_{hi}/{}_aA_{hi} = 1$ or if ${}_bA_{hi} = 0$ for every sample unit the multiple frame estimator is $\hat{X} = {}_{a1}\hat{x} + {}_b\hat{X}$ which is the result occurring when the P + Q weights are 0 + 1, respectively.

We can show that $\hat{X}$ as defined above is unbiased by writing

$$X = \sum_h \sum_i^{N_h} \frac{{}_aN_h}{{}_an_h} \left(1 - \frac{{}_bA_{hi}}{{}_aA_{hi}}\right) {}_ax_{hi} \cdot {}_at_{hi} +$$

$$\sum_h \sum_i^{{}_bN_h} \frac{{}_bN_h}{{}_bn_h} \frac{{}_bx_{hi}}{{}_aA_{hi}} \quad {}_bt_{hi}$$

The value t_{hi} is as defined before for each frame. $E({}_bt_{hi}) = \frac{{}_bn_h}{{}_bN_h}$ because samples are selected independently within strata.

For the area frame $E({}_at_{hi}) = \frac{{}_an_h}{{}_aN_h}$. Since the domain determination is made after the sample is selected, the domain sizes are not known.

Then $E(\hat{X}) = \sum_h^{N_h}\sum (1 - \frac{{}_bA_{hi}}{{}_aA_{hi}})\ {}_ax_{hi} + \sum_h^{{}_aN_{hi}}\sum \frac{{}_bx_{hi}}{{}_aA_{hi}}$

When summing over units in the list (${}_bM_{hi}$ unique units)

$$E(\hat{X}) = \sum_h^{{}_aN_h} (1 - \frac{{}_bA_{hi}}{{}_aA_{hi}})\ {}_aX_{hi} + \sum_h^{{}_bM_h} \frac{{}_bA_{hi}}{{}_aA_{hi}}\ {}_bx_{hi} = X$$

because the weights sum to one across the two frames. The two procedures presented above provide unbiased estimators. The statistical efficiency of the estimators will be explored later. The main difference in the rules is in the complexity of their application. Although all of the procedures result in unbiased estimators, the important point is that they may differ in the bias resulting from the breakdown of the assumptions.

The following rule relies upon a different set of assumptions for defining the overlap between the area and list frames.

Procedure C

We are still referring to Problem 1 as illustrated in Table 1. The assumptions here are:

a. An individual name or the name of a single person on the list represents a unique land operation only associated with that name. More specifically, the name Sam Jones can only represent land operated soley by Sam Jones. It cannot represent land operated jointly by himself and others.

b. If the individual name does not have a unique operation it is considered to be out of business.

When applying these assumptions to Problem 1, we obtain the following results:

a. The parcel of land in the area frame sample operated by the four people mentioned does not overlap with a list frame unit. The operation would be overlap only if a list unit consisted of the four names.

b. There is no duplication in the list frame since each name will only report for land unique to itself. Thus, the estimator does not rely on the ${}_bA_{hi}$ factor.

Procedure C does not rely upon the assumption that every person in a partnership operation will report for the entire operation. Instead, it relies on the assumption that an individual name can only report for individual data. As a result, the amount of overlap between the list and area frames is decreased by Procedure C which then should increase the size of the nonoverlap domain. This is especially true as the list frame becomes more and more out of date - meaning that changes in names of operations or changes in partners will result in fewer matches between the two frames.

Problem 2

This involves the same partnership operation considered above. However, it is complicated by the fact that one of the names is duplicated in the list. (refer to Table 1) Procedures for handling this problem follow.

Procedure A

Since we assume each partner listed in the list frame will report for the entire operation, the land in the area frame overlaps land represented by the list. The factor ${}_bA_{hi} = 3$ because we assume each name selected will identify all of his partners. The important point here is that ${}_bA_{hi}$ equals the total number of list names representing the operation regardless of the fact that some names are duplicates. We also assume the check processes used will identify the name that appears more than once.

Procedure B

Again we assume every individual partner will report for the entire operation. The operation will be assigned to the area and list frames as follows:

a. ${}_aA_{hi} = 4$ because there are 4 members in the partnership.

b. ${}_bA_{hi} = 2$ because there are 2 names on the list that will report for the operation. The duplicated name has the additional factor ${}_bA'_{hi} = 2$

Then the factor for the area frame unit is $(1 - \frac{2}{4}) = \frac{1}{2}$ and the factor for first list name is $\frac{1}{4}$. The duplicate names each will have the factor $\frac{1}{4} \times \frac{1}{2} = \frac{1}{8}$. The sum of weights is 1.0.

Procedure C

Here we assume a single person name can only represent a unique land operation. Therefore, the partnership land operation found in the area frame sample does not overlap the list frame. This also means that the only duplication in the list frame is the name listed twice. Then ${}_bA_{hi} = 2$ for that name. The name(s) on the list can only report for operations unique to that name and not the partnership operation.

The three rules differ in the complexity required to carry out the different assumptions and whether the assumptions apply in practice. A final problem is shown to further illustrate difficulties involved in applying the different procedures.

Problem 3

This involves a parcel of land in the area frame sample that is jointly operated by three people. Two of the three people are linked together as one sample unit in the list frame. Again, we will illustrate how each procedure would apply.

Procedure A

If we assume each individual will report for the entire operation, then we must assume the list name will also report for the operation. There is the risk however, that the list name represents a different operation. With a limited survey time period, decisions must be made quickly. Therefore, by following the assumptions it is determined that the area frame land operation is overlapped by the list frame. There is no duplication in the list frame, therefore ${}_{b}A_{hi} = 1$. One could as easily assume the list name is a different operation, thus the area frame would not overlap the list.

Procedure B

The rule in this case causes confusion in practice because ${}_{a}A_{hi} = 3$ and ${}_{b}A_{hi} = 1$. The weights $(1 - \frac{1}{3})$ and $\frac{1}{3}$ do add to one however. The confusion occurs because two of the names are linked together as one sampling unit.

Procedure C

The name of the area frame operation is not on the list, therefore there is no overlap with the list frame. It is assumed the list frame unit will only report for land unique to a Smith and Brown.

Results

As was stated before, multiple frame estimation requires that the overlap between the sample frames be identified. In other words, the components ${}_{a1}X$ and ${}_{a2}X$ must be accurately determined for the multiple frame estimator to be valid.

The variance estimator

$$VAR\ X = Var\ {}_{a1}X + P^{2}VAr_{a2}X + Q^{2}Var_{b}X + 2PCOV_{a1}X_{a2}X$$

only measures the variability due to random sampling. It gives no measure of the accuracy of the overlap determination. The inaccuracy of the overlap determination falls in the realm of nonsampling errors which are difficult to measure. Since many of the problems associated with overlap determination also affect procedures for handling duplication in the list, additional nonsampling errors can occur.

Each procedure used for the overlap determination relies upon a set of assumptions. Whenever an assumption fails, errors occur. The more complex a set of rules becomes, the more likely it is that inconsistencies will occur. This is especially true if judgement is required in determining if a set of names really do match and that they will report as required by the assumptions.

The procedures illustrated above were each used in an operational multiple frame survey designed to estimate total hogs and pigs on farms. The purpose was to examine the difficulties with applying each procedure and to measure the differences in the estimates and sampling errors resulting from each procedure.

The sample for the survey consisted of about 2,200 farming operations from the area frame sample. The names associated with the 2,200 farming operations were matched with names on a list containing some 80,000+ potential farm operators. Area frame names matching list frame names constituted the overlap domain. The domain determination was done using each of the three procedures.

A sample of 1,600 names was selected from the list frame and also included in the survey. Partnership operations and duplication in the list were processed using each of the three procedures. This allowed us to compute a multiple frame estimate based on each procedure.

Survey estimates based on each procedure and their sampling errors were then computed. The results appear in the following table.

Table 2--Multiple Frame Estimates and Sampling Errors resulting from three procedures for defining overlap between sample frames

Procedure	Multiple frame estimate	Sampling error
A	13.2	.5
B	13.4	.5
C	14.1	.6

Procedures A and B gave similiar results, but then their basic assumptions were also the same. Procedure C differed considerably in the results.

Remember that the three procedures were all applied to the same sample and that unbiased estimators were used.

The sampling error of the difference between any two of the estimates was about .2. This shows that Procedure C resulted in a significantly different estimate from that resulting from A and B.

The larger estimate resulting from Procedure C resulted primarily from an increase in the estimate from the nonoverlap domain. Theoretically, any increase in the nonoverlap domain should be offset by a decrease in the overlap domain and list frame estimate; however, this did not occur. This indicates a problem with a key assumption:

Procedure A & B: Every individual in a partnership will report for the entire partnership and will correctly identify all other partners.

Procedure C: An individual will only report for individual operations.

We can only compare the procedures by evaluating the total error involved, i.e., sampling error plus nonsampling error. The problem is that all three procedures involve some subjectivity. This involves the accuracy with which the respondent can define his operation whether it be an individual or a partnership operation and can be affected considerably by the questionnaire design.

Since procedures A & B resulted in the lower estimates, the assumption that every individual in partnership report for the entire operation may not be met. Procedures A & B also involve more subjectivity and complexity because of the necessity of determining A_{hi} factors for partners and for duplication. Procedure C is less complex and therefore should be easier to implement in an operational survey. However, the assumption for Procedure C may also be failing; that is, an individual may report for more than his individual operation. However, this is doubtful.

Summary

The multiple frame methodology is a powerful survey technique to maximize the efficiency of a survey. However, there are problems associated with determining the overlap between the sampling frames. These problems deserve as much attention as does the development of more efficient estimators. Perhaps the development of new estimators should recognize and measure errors that can occur in the overlap determination because procedures followed in an operational survey can affect estimates an amount greater than what is explained by sampling variability.

References

1/ Hansen, M. H., Hurwitz, W. M., Madow, W. G., "Sample Survey Methods and Theory," John Wiley and Sons, New York, 1953, Vol. 1.

2/ Hartley, H. O., "Multiple Frame Surveys," paper given at Minneapolis meetings of the American Statistical Association, September 1962.

3/ Fuller, Wayne A. and Burmeister, Leon F., "Estimators for Samples Selected From Two Overlapping Frames," proceedings of the Social Science Section of the Montreal Meetings of the American Statistical Association, 1972.

4/ Fellegi, I. P. and Sunter, A. B. (1969), "A Theory for Record Linkage," Journal American Statistical Association 64, pp. 1183-1210.

5/ Gurney, Margaret and Gonzalez, Maria Elena, "Estimates for Samples From Frames Where Some Units Have Multiple Listings." Proceedings of the Montreal Meetings of the American Statistical Association, 1972.

6/ Rao, J. N. K., "Some Non-Response Sampling Theory When the Frame Contains an Unknown Amount of Duplication," Journal of the American Statistical Association, March 1968 (87-90)