FIDDLING WITH AREA FRAME INFORMATION

IN LIST DEVELOPMENT AND MAINTENANCE

by

Fred Vogel

Dwight Rockwell

gdh

CONTENTS

FIDDLING WITH AREA FRAME INFORMATION IN LIST DEVELOPMENT AND MAINTENANCE

INTRODUCTION

An important priority of the Statistical Reporting Service (SRS) is the development of a general purpose list sampling frame. The uses and benefits of such a list are obvious and documented elsewhere. Given this intensified emphasis on list building, there is an ever-present temptation to use any and all information available, including that obtained in connection with area frame sample surveys.

Regardless how much information and effort goes into building and maintaining a general purpose list, it will never be complete on a current basis. Thus, the unique contribution of an area frame to any multiple frame estimator is the estimation of incompleteness in the list. This information is available from no other source. Therefore, we must be ever vigilant to avoid any action, no matter how trifling the action may appear, that might undermine the area frame's ability to estimate list incompleteness.

During recent years, members of the Sampling Studies Section, including the authors, have visited several states to collect data for research projects related to multiple frame sample surveys. Through these experiences we have learned that tracing the effect of using area frame information to improve the list is often not easy in practice. Conclusions based on intuition rather than a thorough understanding of the relationship between the area and list frames and of how multiple frame estimators are constructed can easily be wrong.

The purpose of this paper is to examine specific reasons for the detrimental effect of using area frame information to improve the list. A series of questions are proposed for discussion. First, characteristics of SRS list and area frame sample selection and rotation procedures that are pertinent to this discussion will be outlined.

SRS SAMPLING AND ROTATION PROCEDURES

Each time an SRS list frame probability sample is drawn, each sample unit (name) on the list has a known probability of selection, and if the probability is not greater than zero for a given unit, the unit is not a member of the list for this particular survey. Currently, when a new list frame probability sample is drawn following a list update, it is done independently of any previous sample selection. That is, the probability of selection for each sample unit does not depend on whether or not it was selected for any previous survey. If in the future, list units are selected with probabilities dependent upon previous sample selections to reduce respondent burden, estimating procedures will require corresponding modification.

The SRS area frame is constructed by stratifying and then subdividing all land in each State into sampling units (segments) with a known probability of selection. Since every parcel of land has a chance to be selected, with no duplication or omission, the area frame is said to be "complete". That is, the entire population of interest is wholly contained in this sampling frame. A replicated random sample is selected without replacement. For each year's survey, random subsets of the previous year's sample segments (replications) are rotated out and a replacement set is rotated in. To reduce respondent burden, rotated-out segments are not permitted to be re-selected (zero probability) for five years.

DISCUSSION

1. Why cannot the name of a nonoverlap (NOL) tract operator found in a previous year's non-rotated area frame sample segment be added to the list when the name does not appear in an independent list source?

Each time the list is updated a number of names from a source(s), not including the area frame, are added to the list. It is to be expected that a few of these new names will be of NOL tract operators in non-rotated segments. For instance, if the expansion factor for a NOL operator was 200 and the name of this operator were found in an independent source, one would expect to find about 200 additional names in this same source. Since we are speaking in terms of expected values, this precise relationship (1 to 200) may never actually occur.

Following a list update, the remaining NOL tract operators will represent the incompleteness remaining. If at this point the name of a NOL tract operator is arbitrarily added to the list, the list certainly is improved. However, this NOL tract operator can no longer represent other operators whose names have not been added to the list. The downward bias resulting from such an action is a function of the area frame sample expansion factor.

2. Why cannot the name of a NOL tract operator found in a previous year's rotated-out area frame sample segment be added to the list when the name does not appear in an independent list source?

The most important reason this action should not be taken stems from the procedure by which new segments are rotated into the area frame sample. To reduce respondent burden the previous year's segments that are rotated out are given no chance of being rotated in.

However, each year's area frame sample estimates for the total nonoverlap domain, including those rotated out. Therefore, if the names of NOL tract operators who have rotated out were added to the list, the current sample segments would not reflect the removal of these names from the nonoverlap domain causing nonoverlap to be overstated. The multiple frame estimator would be biased upward.

3. *Why cannot the names of NOL tract operators in segments of an old area frame sample be added to the list when a completely new, independent area frame sample is drawn?*

If we were certain the old frame sample would never be used, there is no reason for this action not to be taken. Furthermore, the list would be improved. Since use of the new sample will provide relief in response burden to old frame reporters and provide more efficient estimates, use of the old frame in the future is not currently anticipated. However, if there is need to use the old frame it will not provide an unbiased estimate of list incompleteness if NOL operators have been added to the list.

4. *Why cannot names of NOL tract operators be used as a source of names for nonprobability mail surveys?*

Names for both nonprobability and probability surveys are expected to be included in the same general purpose list sampling frame. Although names of NOL tract operators may be intended for nonprobability surveys only and are coded accordingly, including these names on the list could result in these tracts becoming erroneously classified as overlap for subsequent probability surveys.

5. *Under what condition can the name of a NOL tract operator not be added to the list, even though the name is found in an independent source?*

Often the names of a few NOL tract operators become very familiar to a survey statistician because an unusually large number of livestock are involved, or the tract has been in the area sample for several years. If one of these names appears in a farm magazine, local newspaper, etc., it probably will be recognized. If all names appearing in this source are not added to the list on a regular basis, then NOL tract opertors' names should likewise not be added. To do so would artificially reduce the number of NOL tracts and the list incompleteness would be understated.

6. *Occasionally an overlap tract operator will report a number of livestock for an area frame survey that is substantially different from the list control number. Why cannot this information obtained through the area frame be used to update the list control number?*

This action could result in adding a name to the extreme operator (E.O.) list. The incompleteness of the E.O. list would be understated for subsequent June and December Enumerative Surveys.

It would be permissable to change list control data using area frame information as long as it did not result in adding a name to the E.O. list. However, this decision would be based on current E.O. cutoffs (minimum control number for E.O. classification). These specifications have changed substantially over the years in many states as survey procedures have changed. Since future cutoff changes can be expected, the ultimate effect of updating list control data based on area frame information cannot be known with certainty.

7. *Each time a list is updated, a number of names is usually dropped because according to usual list sources, they are believed to represent persons who no longer operate a farm. What action should be taken if a name to be dropped is recongnized as that of a tract operator.*

If based on all list information available, it has been decided to drop a name, that decision should be adhered to even though the name is of a tract operator. The decision to drop or retain a list name must be made independently of any area frame information. To do otherwise would have the same effect as adding a NOL tract operator's name to the list.

8. *During the process of classifying area frame tracts as overlap or nonoverlap, one frequently finds a tract operator's name on the list, but the match is not absolutely perfect. It is not unusual to find a different name spelling or a different or less complete address. A farm name will sometimes be provided on the tract questionnaire, but not on the list source. Many such tracts are classified as overlap because the differences are considered minor or because further checking, including possibly a re-contact with the tract operator, shows the two names are, practically without a doubt, for the same person. Why under these conditions should the name on the list not be changed to correspond exactly with that on the tract questionnaire?*

The two activities, overlap determination and list building, must be regarded as two separate operations. Overlap determination must be conducted for each survey based on current information available. Sometimes a tract will be classified overlap even though the match with the list is not exact. This decision may be made because a check of telephone books, county directories or

even a recontact using one or both of the names involved leads one to conclude these two names are for the same person. Recognizing that errors in judgement may occur, the best choice must be made. This decision can be reviewed during subsequent surveys as new information becomes available.

If after classifying a tract overlap, a name or address is changed or made more complete on the list to match exactly that of a tract operator, the tract classification probably becomes permanent. If an error was committed, there is a good chance it will never be corrected. The resulting bias in a current survey estimate will also become a bias in subsequent survey estimates. Also, by making a change in the list names, we may affect how the farmer will report if selected from the list and how we may edit his list report. Reports from farmers who are not in the area frame sample would not be similarly affected.

SUMMARY

This discussion is not intended to cover all possible questions that arise concerning the use of area frame information to improve the list. Hopefully the more important cases have been covered.

Sometimes the reasons for not using area frame information to improve the list are not obvious. However, in each case we should ask ourselves the following questions:

1. Does this action affect how the area frame tract will be classified in the future (i.e., overlap or nonoverlap)?

2. Will this action affect how the farmer will report if his name is selected from the list?

3. Will this action affect how a report will be edited if this name is selected from the list?

4. Do we really know the answer to the above? Is using area frame information to change the list worth the risk of impairing the area frame's ability to estimate for incompleteness in the list?

With the current emphasis on the development and maintenance of a general purpose list sampling frame, SRS statisticians must be particularly concious of potential misuses of area frame data. The validity of our multiple frame sample estimates and the independence between the two frames must be preserved. The only way we can be assured this independence is maintained is to never use area frame information in list development and maintenance.

The area frame sample is still and will be our "bread and butter" for providing an essential foundation for the SRS survey program. The list frame is assuming its proper role, but when the two are combined into a multiple frame sample design, it is mandatory that the basic sample units are kept separate and independent.

In a practical sense, this is not easy to do because the same individuals in a State Statistical Office will be working with both frames. Therefore, it is important to understand the basic concepts to be able to maintain independence between the frames.