RECENT AND FUTURE ACTIVITIES TO IMPROVE SAMPLING FRAMES FOR AGRICULTURE

Ron Fecso, Robert D. Tortora, and Frederic A. Vogel
U. S. Department of Agriculture

INTRODUCTION

Information about current and future supplies of agricultural commodities is needed by farmers, ranchers, and agri-business firms for marketing, planning, and decision making. This information is also necessary for policy decisions concerning government programs affecting the agricultural economy in specific ways and the U. S. and global economies in more general ways. To meet these needs the Statistical Reporting Service of the U.S. Department of Agriculture (SRS) publishes about 300 national and 9,000 state reports each year. These reports cover a broad range of agriculture including about 120 crops and 45 livestock items (23).

Agriculture in the U. S. is a business that consists of 2.4 million farms that show tremendous diversity in size as well as the types of products produced. This diversity has many implications about the sampling methodology necessary for an efficient survey program.

Farms vary widely in size as measured by total value of production. One third of the farms account for over 90 percent of the total value of production. (A farm is a place producing $1,000 or more of agricultural products). One percent of the farms account for a third of the total sales. On the other hand, the farms differ considerably in what is produced. Only 10 percent of the farms account for three-fourths of the corn acres. Less than three percent of the farms produce crops such as peanuts, cotton, or rice.

Therefore, agriculture can be characterized as a population that first varies tremendously in size with a large number of small operations and a small number of extremely large operations. Second, the overall population of farms consists of many subgroups that really constitute rare items when considered in a sampling sense. This diversity in size and the need to sample and survey for rare items has led to the development and use of multiple frame sampling procedures relying upon area and list sampling frames.

The area and list sampling frames each have strengths and weaknesses as they relate to the characteristics of the U. S. farm population. This paper details some of these strengths and weaknesses, outlines current and proposed research directions and discusses policy issues regarding agricultural sampling frames.

AREA FRAME MERITS

The area frame is complete in the sense that all farms and land have a known probability of selection. The frame is suitable for general purpose type surveys that cover a wide spectrum of crop and livestock items. It can also be used for economic type surveys where the reporting unit is either a farm headquarters or a farm household. Although the initial investment in developing the frame can be considerable, the life span of an area frame can be long, which is beneficial for longitudinal type surveys.

A weakness of the area sampling frame is that is is inefficient for commodities represented by extremely large farms or commodities that are rare in that they are only produced on a few farms.

Some items such as cattle, which are produced on a large proportion of the farms, are also characterized by tremendous variability by size of operation. For example, five percent of the farms account for two-thirds of the total cattle inventory. Therefore, the main concern when designing a sample for items such as cattle is to reduce the variability caused by the extremely large operations.

Rice is typical of an item that is produced on farms showing less variability in size. However, only .5 percent of the farms produce rice. This means that a general purpose sample of area frame segments would only yield about one in 200 farms actually reporting rice unless some crop specific stratification was employed. In either case, the main source of sampling variability is caused by the rarity of the item which inflates the sampling variance.

LIST FRAME MERITS

For these reasons, SRS has also relied upon the use of list frames to supplement the area frame in its survey program.

Lists of farm operators have been used in the Agricultural Statistics program almost from its inception. In 1882 part-time statisticians were appointed to develop and maintain groups of voluntary crop reporters to provide current information about agriculture. In 1892, 125,000 farm operators were furnishing survey data for annual estimates. The generalized structure of agriculture through the middle of the 20th century allowed the Department to rely upon general purpose lists for its estimating program. By the early 1960's however, agriculture was becoming more specialized and a number of extremely large operations began emerging.

At the same time, SRS began shifting its survey program away from the general purpose non-probability surveys to the area frame probability survey. As mentioned above, the emergence of large farms and the increasing specialization of agriculture led the Agency in a search for procedures to supplement the area frame. Research by Hartley (17) led to the implementation of multiple frame sampling which called for the joint use of both area and list sample frames.

The main strength of a list frame depends upon how it is constructed, but should include:

(a) It should either be complete for the item being estimated, or be nearly complete for the size or type of farms to be represented by the list frame in a multiple frame survey.

(b) Measures of size should be available for each item of interest to indicate its presence and the relative size for efficient survey design purposes.

Because of the dynamic nature of agriculture, the task of compiling a complete list is cost prohibitive. Therefore, the main strength of a list frame is to supplement the area frame's weaknesses - that is for rare items and for items with extreme variability.

AREA FRAME OVERVIEW

Two aspects of area frame sampling in agriculture are now discussed. They are:

1) developments in the construction and sample design since development of the master sample of agriculture (20), and

2) prospects for improving the construction and maintenance of area sampling frames in the future (11).

Since 1967 SRS has been using area frame sampling in all 48 conterminous states in a system of surveys for obtaining information on crop acreage, livestock numbers, grain production and stocks, costs of production, farm expenditures and other agricultural items and as a basis for subsampling for crop yield and other speciality surveys (23). Changes in the area frame design were slowly adopted over the 35 year period from 1940 to 1975. These changes represented a switch from the master sample concept to a frame which utilized land-use stratification. The master sample frames were constructed on county highway maps with minor civil divisions and sample units delineated on these maps. Each sample unit contained about four farms while crop reporting districts within each state were used to provide geographic stratification. The changes included a refined stratification process and the introduction of replicated sampling. Until the mid-seventies, the area frame construction and maintenance process can be characterized as being essentially the same paper and pencil operation, using the same types of materials, as used for the master sample of agriculture. After 1975, the impact of new technologies affected the area frame construction process. The computer was incorporated at several places in the process, from measuring the land area of the frame and selecting the sample to providing quality control for the construction process (9)(10). Table 1 summarizes the significant chronological events in area frame sampling for agriculture. Notice that while the changes made in the sixties and early seventies are primarily related to sampling methods, such as the new stratification by land use and the introduction of interpenetrating sampling, the changes that began in the mid-seventies represent the application of new technology to area frame construction, as exhibited by the uses of the computer and the availability of satellite imagery.

In 1978, SRS replaced the last master sample frame with frames stratified according to land use (18). For comparative purposes some of the characteristics of the master sample frame and the current SRS area frames are given in Table 2.

There was about a sixty percent drop in the total number of U.S. farms between 1945 and 1984. The new

TABLE 1 — Significant events in area frame construction

YEAR	EVENT
1938	Iowa State University (ISU) begins construction of area frames for the master sample of agriculture
1954	The Statistical Reporting Service (SRS) begins investigating the use of area frame sampling
1962	Land-use stratification is introduced in SRS area frames
1967	All 48 conterminous states have area frames
1973	Interpenetrating sample designs introduced
1976	Computer selection of area frame samples
1978	The last state having a master sample is replaced by a frame having land-use stratification
1979	ISU discontinues area frame construction
	Digitized area frame files created for each new frame (Manual planimetering discontinued)
	Satellite imagery used in stratification
	Crop-specific stratification introduced
	Initial development of the Area Frame Analysis Package
	Area frame development for remotely sensed sampling in foreign countries begins
1980	Minicomputer used for quality control of area frame construction procedures
1981	Area frame data base developed
1982	Use of National High Altitude Aerial Photography initiated
1984	Automated area frame management system developed

frame differs in that the number of sample segments has decreased by over seventy-five percent and resident farm operators have decreased by about ninety-five percent.

The land-use frames are constructed on mosaics of aerial photographs and then transferred to county highway maps in order to accurately measure land areas. Table 3 shows the flow of the construction process and denotes the new use of computer intervention in this process. In general, the urban and rural places strata of the master sample are still being used by SRS in its land-use frames. However, the Open Country stratum has been further subdivided to obtain improved sampling efficiency. To begin land-use

Table 2 — The master sample and land use area frames (1945 compared to 1984)

Frame characteristic	Master sample	Land use area frame
US Number of farms	6 million	2.3 million
Sample Size	67,000 segments	16,000 segments
Resident Farm Operators in Sample	300,000	16,000
Measure of Size	Indicated Number of Farms	Area of sample unit (segment)
Stratification	By Crop Reporting District:	
	Urban Places	Land Use
	Rural Places	Potential Urban
	Open Country	Crop Specific

Table 3 -- Area frame construction: An increasingly automated process.

Process	Description
Stratification	A manual process of delineating homogeneous blocks of land, or primary sampling units (psu's), on aerial photographs and county highway maps
Digitization	The area of the psu's is measured through use of a microcomputer and digitizing tablet.
County Level Edits	Data transferred to a minicomputer to perform consistency checks at the county level
Digital Area Frame Finalized	Main-frame computer used to: a. edit at county and state level b. obtain measures of size for pps first-stage sample selection c. select first stage units d. archive the area frame
Second-Stage Selection of Segments	A manual process of defining ultimate sample units (segments) on aerial photography

stratification, blocks of similar areas of land are identifed within each county (counties are used as a tool to manage the work flow of the area frame construction process) and classified into one of the following strata: 1) intensely cultivated areas where a significant portion of the land is under cultivation, 2) extensively cultivated areas used primarily for grazing and producing livestock, 3) agri-urban areas around cities, 4) urban areas, and 5) nonagricultural land such as parks and military reservations. Of course, each of the above strata can be further subdivided to take advantage of geographic differences or agricultural specialization that may exist within a particular state. Table 4 illustrates the strata that are currently being used in Idaho.

After stratification the primary sampling units are defined on the photo mosaics. The average size of a psu varies by stratum. For agricultural strata, a psu contains an average of 8 sampling units or segments. During the construction of the primary sampling units, the main emphasis is to delineate units that can be further subdivided into homogeneous segments using observable boundaries that can be easily found by an enumerator during data collection. The area of the primary sampling units is obtained by digitizing the county highway maps. A statistical package computer program is then used to plot each county to ensure that each psu has been digitized and assigned to its proper stratum. After a state is completely digitized, a sample of first stage units is selected. Each selected psu is then further subdivided on the photo mosaic into segments and one segment is selected at random. Except for unusally large segments in rangeland areas, or for some segments in large cities, photo enlargements are provided for enumeration.

The following section gives a more detailed discussion of the developments in area frame construction and sampling since the Master Sample. The last section outlines the prospects for these processes in the future.

AREA FRAME DEVELOPMENTS

Area frame construction is a major undertaking which must be considered as a long-term investment. The

Table 4 -- Stratum Definitions for an area sample frame, Idaho.

STRATUM	DEFINITION
10	Dryland Grains--small grains, primarily wheat and barley, 33 percent or more cultivated. This stratum will be found primarily starting in Idaho county and Northward and in the Southeastern counties of Fremont, Madison, Teton, Bonneville, Caribou, Bannock, Powers, Cassia, Oneida, Franklin and Bear Lake.
13	General Crops--50 percent or more cultivated land outside the Snake River Basin that is not dryland grain. Majority of cultivation expected to be irrigated small grains.
15	General Crops--50 percent or more cultivated along the Snake River, all irrigated, intensively cultivated land in Canyon, Ada, Owyhee, Elmore, Gooding, Twin Falls, Lincoln, Jerome, Mindoka, Cassia, Power, Bannock, Caribou, Bingham, Bonneville, Tenton, Madison, Jefferson, Fremont, Clark, and Buttee should be in this stratum. This stratum should contain practically all of the potatoes and sugar beets. NOTE: A county might have strata 10 and 13 or 10 and 15. It is not possible to have 13 and 15 in same county.
20	General Crops--15 to 49 percent cultivated. Includes extensively cultivated land outside the Snake River areas that is not in dryland grains.
22	Dryland Grains--15 to 33 percent cultivated. Extensively cultivated land used in conjunction with stratum 10. (Maybe collapsed with stratum 20 if area insufficient in size to justify a separate stratum.)
25	General Crops--15 to 49 percent cultivated used in conjuction with stratum 15.
31	Agri-urban--More than 20 dwellings per square mile, residential mixed with agricultural.
32	Residential Commercial--More than 20 dwellings per square mile, no agriculture present.
33	Resort--More than 20 dwellings per square mile. May be collapsed with stratum 31 if size of land area insufficient to justify a separate stratum.
40	Rangeland and Pasture--Less than 15 percent cultivated. Includes both public and private range. Woodland and forest would also be included.
50	Nonagricultural Land--Land not used for agricultural purposes and usually documented by law or other regulation. This stratum included such land uses as airports, wildlife refuges, military installations, National and State parks and so forth.
62	Water Bodies--1 square mile or larger

efficiency of the frame over time will be a direct result of the frame construction procedures and the sample design chosen. Recognizing the importance of these decisions, SRS has maintained an ongoing research effort to improve area frame sampling. This section will outline the major changes to the SRS area frame made since the master sample was used.

Through the past few decades, research and operational experience have resulted in an evolution of area frame construction. The research in area frame sampling has been directed toward the search for cost saving techniques and methods which will improve the efficiency of the estimators.

Stratification

One of the first major changes to the master sample concepts was stratification by land-use. Starting in the early 1960's, master sample area frames were replaced on a state-by-state basis by area frames which incorporated land use stratification. Generally, six land-use strata based on the amount of land cultivated were used. These general strata were intensive agriculture, extensive agriculture, cities and towns, range, nonagriculture, and water. As experience was gained, some of these strata definitions were further subdivided to create strata which would solve specific enumeration problems such as too many agriculture tracts in a segment or overly dense residential development (2).

By 1978, all states had area frames with a form of land-use stratification. These frames continue to be updated at the rate of 2 or 3 per year. The area frames do not become out-of-date in terms of population coverage, but the efficiency does deteriorate over time. Land subdivision results in increased enumeration problems, boundary changes present a potential bias and the land-use within strata changes. Experience has shown that each state has a unique set of enumeration problems, materials available for frame construction, and estimation requirements and priorities. Based on this experience, the thrust of research since 1979 has been

toward the development of a timely, yet thorough, analysis of the requirements and best methods to use for each state which will have a new frame constructed (10). The main outgrowth of this effort has been the use of crop-specific stratification in states which have concentrations of important crops which can be identified with the available materials. Examples of crop-specific strata include fruits and vegetables in California (new frame in 1979), dryland grains in Washington and Oregon (1980) and Idaho (1982) and rice, cotton, wheat and peanuts in Texas (1982).

Because the SRS area frame is used to collect multiple data items, there has been much debate over what is the most effective use of stratification (2)(8)(9)(16)(19). Stratification for more than a few specific commodities is difficult and as a result of the chosen sample allocation could reduce the efficiency for other commodities. Recent experiences with crop specific stratification have exhibited desirable results. Creating certain crop specific strata results in more efficient estimation of the specified crop while other crops are less of a rare item in the remaining general strata and thus more precisely estimated.

Replicated Sampling

Frames constructed since 1974 are sampled using a replicated design (20). Replicated sampling is characterized by the selection of several independent samples from the frame. It was initiated to facilitate the rotation of sampling units in order to limit individual respondent burden. Other advantages of replicated sampling include the use of subsets of the replicates for special sampling purposes such as one-time surveys or nonsampling error studies and the ease of variance computation (useful especially in underdeveloped nations and for special surveys).

Replicated sampling, as done by SRS, utilizes a form of substratification called "paper stratification" which essentially is a geographic substratification of each state (14). The first step in paper stratification is to determine a meaningful ordering of the psu's in each stratum. To determine this ordering, a cluster analysis of the agricultural estimates for each county is examined to determine "similar" agricultural areas. The result of this analysis is an ordering of the counties such that, to the extent possible, similar counties are in sequence through the ordering. Since all psu's are identified by county, the frame can be sorted to arrange psu's in this county order in each stratum. Once ordered, the stratum is divided into several pieces (paper strata) each with an equal number of sampling units, except the last piece when the stratum size is not exactly divisable by the number of paper strata. Strata with few sampling units usually have 2 or 3 paper strata, while large strata may have from 10 to 20 paper strata.

A replicate in the SRS design is defined as a simple random sample of one sampling unit (segment) from each paper stratum in a land-use stratum. The paper strata thus serve much the same purpose as systematic sampling in dispersing the sample throughout the population, but in essence they are a form of commodity specific stratification which contributes to the efficiency of the estimates.

Materials

An important problem in SRS area frame construction has been the age of frame construction materials. The technological advances in agriculture, especially in irrigation, over the past 20 years have vastly expanded the cultivated areas. Cropland expansion and urban development wreak havoc on the efficiency of the land-use stratification and create problems for enumerators when the photography is old. Eventually, enough gain in efficiency can be realized from restratification to justify the cost of new frame construction. In order to get the most gain from a new frame, current materials are essential. Prior to 1979, stratification was done using Agricultural Stabilization and Conservation Service (ASCS) photo mosaics. These materials were often 20 years old in some areas and rarely less than 3 years old. Also, in some areas no coverage was available and United States Geological Survey (USGS) topographic maps had to be used. To overcome some of these problems by providing more recent data for frame stratification, Landsat satellite imagery was incorporated into the stratification process in 1979.

Easily identified segment boundaries are a requirement for effective enumeration. Boundary quality received an assist with a new program for high altitude color infrared photo coverage, the National High Altitude Photography (NHAP) program, done as a cooperative venture by various government agencies. The goal of this project, which began in 1980, was to have complete coverage of the continental U.S. which is never more than three years old. As this material becomes available it is being used in the construction of new frames.

Sample Allocation

Crucial to the successful use of a new frame is the allocation of the sample to the strata. Research continues on methods to allocate the most efficient sample during the first year the frame is used. Recent advances include the development of estimators for the optimum allocation of a replicated design, the post-stratified use of prior survey data to measure stratum variances in the new frame, and the use of data collection times for each strutum so that cost data is incorporated in the allocation formulation (9).

Quality Control

Quality control of the SRS area frame is achieved in several ways and the development of improved controls is a continuous effort. Prior to 1978, most quality control was a result of a post-survey review of the data collected from the sample units. In 1978, quality control procedures for the random selection of secondary sampling units and the frame construction review process were initiated. In 1979, a post-survey analysis package was developed which helped point out statistically inefficient frame construction techniques as well as construction errors (10). By 1980 a minicomputer system was being used to do quality control on many processes prior to the sample selection.

Another Area Frame Design

The Sampling Frame Development Section worked on a cooperative agreement with NASA to construct area frames for Georgia, North Carolina, and portions of Argentina and Brazil (7). These frames were to be used for crop area estimation using remote sensing methods in the AgRISTARS program and are thus considerably different from our usual area frames. During the construction of these frames from 1979 to 1982, SRS developed methods and gained considerable experience using remote sensing techniques and associated computer applications. This experience resulted in the application of the relevant technology to domestic area frame construction and use.

The use of Landsat as a source of auxiliary information for alternative sample designs and to improve area frame estimates of specialty crops is an active research item. Another AgRISTARS outgrowth was the development of a system of microcomputers with digitizers and a minicomputer to enhance the quality control of the frame construction process and to reduce the costs of both manual and automated processes.

New Frame Analysis

Before construction of a new frame begins, a considerable amount of information is assembled and analyzed to help in the decision on how to achieve better estimates for a fixed cost of sampling. This information includes obtaining the available Landsat imagery, determining the age of the aerial photos to be used in stratification, evaluating the impact of the June Enumerative Survey (JES) estimates on state and national precision, reviewing county estimates, gathering data on urban development and changes in land usage, and analyzing prior years' JES data. The information is used to determine the type of stratification which would be most efficient in the particular state. When the stratification is complete, the prior years' JES segments are located on the new frame and post-stratified in order to provide a more analytical estimate of the stratum variances and a better initial allocation of the sample. An area frame database is being developed to provide much of this information. As each year's data is added to the database more information will be available to determine which states should have a new frame.

After the first use of the new frame, the JES survey data is processed through the Area Frame Analysis Package (11). This analysis package provides graphical and statistical information to allow a detailed analysis of the frame construction and the sources of variation. Often the analysis uncovers nonsampling errors, improved allocations and design or construction alternatives which could be useful in future frame construction.

AREA FRAME PROSPECTS

SRS's revitalized interest in area frame research began in 1978. The current research staff is working on several projects which will have both short and long range impacts on area frame construction and use. Some of the projects will be outlined in this section.

Landsat

The multispectral data and associated imagery from the Landsat satellite possess the greatest potential for improvements in area frame sampling (5)(9)(16). If processing costs decline and classification of the multispectral data into land cover classes improves, there are several major changes that could take place.

The most immediate application of Landsat technology which can be developed is in the second stage of sampling. Here a psu is divided into a predetermined number of sampling units with the constraints of good boundaries and keeping the sizes nearly equal. Landsat data for the psu could be used to help achieve a division which is homogeneous with respect to the various crop acreages estimated. This process could reduce a substantial portion of the variance of crop acreage estimates from area frame surveys (9).

Research being done using Landsat to detect land use changes presents several possible uses. Auxiliary data estimated for the primary and secondary sample units can be used in regression estimators of crop acreage. These estimators should be more precise than the direct expansion estimates, but problems in acquiring current Landsat data impact the timeliness and cost. An alternative approach to regression estimation would be to use this Landsat data for stratification. Timeliness, efficiency, and cost may all be improved by using the measures of change in auxiliary data for psu's to improve stratification and sample allocation. More recent Landsat data could be used for post-stratification. In essence this would create a geo-reference file which would enhance the efficiency of multidisciplinary and specialty surveys, and estimation of rare items and small area estimation.

Geographic Information System

With the digitization of psu's the possibility of developing a Geographic Information System (GIS) for area frames approaches reality. Together with the addition of auxiliary information created by the digital processing of Landsat information, a GIS will allow the area frame to be operationally maintained more like a list frame. The registration of psu boundaries to a geographic reference system, such as latitude-longitude coordinates, can allow for the automatic updating of each psu as land use changes, e.g. by the encroachment of urbanization or the change of rangeland to irrigated cropland, to reflect this new ancillary information that can be used for restratification. As the psu's are updated, the impact of their changes can be evaluated. When there is sufficient change to the auxiliary data, such that the sample design changes would improve the efficiency of the estimators, the restratification process can be started.

When this potential for the GIS is realized, a significant portion of the construction process can be moved from the manual to the automated mode. Referring to Table 3, we see that the entire process up to segment selection could be done, almost entirely, by machine. Of course, manual intervention will never be completely eliminated from this part of the construction process, but the time needed will be significantly reduced.

Stratification Methods

A specialty area frame was developed in Michigan to estimate dry bean acreage (8). This frame can be used to assess various alternative methods of stratification. Included in the possible studies are: the impact of different psu sizes on cost of construction, optimal strata definitions, substratification, and frame update strategies.

Frame Rotation

Under our old procedures, the development of a new area frame often results in the under-use of many segments. These segments were those in the old frame which are not in the sample for a full five year rotation cycle as well as those in the new frame which are rotated out during the first few years. For example, from 1979-1983, an average of 1050 old frame segments out of about 15,000 sampled nationally were abandoned each year, while during the 1979-1982 surveys an average of about 975 new frame segments took their place. Since 80 percent of each group does not receive full utilization, the concept of rotating into a new frame has the potential for considerable cost savings as well as other benefits.

"Rotating into a new frame" is best described as a combined estimator using the replicated design in the old and new frame. Considering any direct expansion estimate, let

O_i = the estimate from the old frame for year i

N_i = the estimate from the new frame for year i

then the combined (rotation) estimate for year i is

$$C_i = W_i O_i + (1-W_i) N_i.$$

where O_i and N_i are independent estimates which are unbiased to the extent of the frame being used. W_i might be chosen as the proportion of remaining segments in the old frame so each year W_i would decrease by two-tenths. Considering that the new frame should be more efficient, W_i might be chosen smaller than this rate.

Other potential benefits from a rotation estimator are workload reduction in sample selection and workload evenness in the data collection efforts.

Finally, frame updates would be facilitated by a rotation estimator. Whenever improved auxiliary information is obtained, a mechanized update can be implemented. If frame errors are found or the first year allocation is inefficient, remedies can be quick and easy through the replacement of only the small, new-sample rather than a full new-frame sample.

Recently, a procedure for re-use of segments which were rotated out of use during the first few years of using a new frame was developed. Some of the cost savings mentioned above have been captured by this new technique.

In the long run, one can perceive the frame construction task to be a more professionally challenging task for both statisticians and cartographers. The heavy manual work will be replaced with the maintenance of a geographic information system, combining the talents of our remote sensing research, cartographic training and sampling theory.

LIST FRAME

A list frame for agricultural purposes is a list of farm operators. A properly constructed list should contain names, addresses, and measures of size for necessary survey items. In addition, the list should be complete for each item being surveyed and should be free of duplication. Since the process of constructing a list frame initially involves assembling lists from a variety of sources, identifying duplication is a major problem. The match process of identifying duplication using computer technology is called record linkage. The statistical decision model used in the linkage process relies upon the frequency of occurrence of names, address, and other information. The underlying theory for the model used by the Statistical Reporting Service was developed by Fellegi and Sunter (12). By using statistical decision models and other match procedures, a determination is made for each record whether or not that record should link with other records. In linkage, two probability values (threshold values) are used to assign records to one of three groups.

(a) Non-linked records

(b) Probable linked records

(c) Definite linked records

The threshold values are used to separate non-linked records from definite links. All probable linked records are manually reviewed and resolved before sampling takes place. The manual resolution is a difficult, time consuming task. Considerably more research is needed to determine the appropriate location of the "thresholds" and their impact on the subsequent sample design. One alternative would be to allow each "probable link" to remain in the frame and let its probability of selection for a given survey be weighted by its linkage probability.

The primary advantage of a list frame is that if good measures of size are available, stratification can be used to reduce overall sample sizes. In addition, data collection is less costly because data can be collected by mail and telephone.

A problem with a list frame relates to the matter of duplication that remains in the frame and is not detected until the sample has been selected. One solution is presented by Gurney and Gonzales (15) where the number of times a given operation is duplicated is not known. Another method has been developed by Rao (22) for the case where the number of times an operation can be selected from the frame is known.

In practice, an attempt is made to 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.

Another disadvantage of a list frame is that it is usually incomplete and is constantly changing. Not only does

the content of the frame change and names enter and exit agriculture, but the operations show considerable change in their structure and size from year to year. It has been found that about 20 percent of the records in a list frame will change from year to year. Therefore, it is important that savings resulting from sampling and collection efficiencies associated with a list frame exceed the frame maintenance costs.

MULTIPLE FRAMES

The primary reason for using multiple frame sampling procedures is to capture the strengths of the area and list frames. The list frame, while incomplete, can be efficiently sampled for rare and variable items. The area frame is a complete frame, but is inefficient for rare items and items that are extremely variable in size. Therefore, when multiple frame sampling is used, the area frame is primarily used to estimate for the incompleteness of the list frame.

Multiple frame surveys are subject to all operational problems that plague single frame surveys. By their very design, problems unique to multiple frame surveys also occur. These problems arise from 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. That is, 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 or not it could have been selected from another frame also being used. The available theory does not tell how this determination is to be made - it only gives alternative estimators to use once the determination is made.

Two items need to be 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 alone is:

$$\hat{X}_{area} = \sum_h \frac{N_h}{n_h} X_h$$

Where (N_h/n_h) is the reciprocal of the probability of selecting a sample unit in the area frame and X_h is the sample total in the h^{th} stratum. The area frame estimator can also be written as:

$$\hat{X}_{area} = \hat{X}_{nol} + \hat{X}_{ol}$$

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

A multiple frame estimator first presented by Hartley (17) is:

$$\hat{X} = \hat{X}_{nol} + P\hat{X}_{ol} + Q\hat{X}_{l}$$

where $\hat{X}_l$ 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 simpler 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 two domains.

A difficult operational problem associated with multiple frame surveys is the need to divide the area frame into two domains.

If costs were 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 found in area segments against the list frame.

This is probably the most difficult factor involved in a multiple frame survey. Errors in this determination are not considered in the estimation phase, thus they fall into the area of non-sampling errors. The name matching operation can be completed manually or by a method of record matching as described above. Whichever procedure is used requires certain decision logic about what is a match and what is a non-match.

The sampling efficiencies to be gained through multiple frame sampling are illustrated in Table 5. Especially note the gains from a list sample of 450 names of potato producers. The area frame sample would have to have been increased by a factor of 9 to achieve the same sampling precision provided by the multiple frame estimate. The gains in this case are especially clear if the size of the area frame sample is adequate for other items being estimated.

Several factors need to be evaluated when considering the use of multiple frame sampling. For example, with agricultural surveys it is generally accepted that an area frame is necessary to provide complete coverage of the population. Therefore, the costs associated with area frame development can be considered to be fixed. The size of the area frame sample depends upon two factors:

(a) The size needed to adequately estimate for items for which it is reasonably effecient.

(b) The size needed to estimate for the incompleteness of the list frame if a multiple frame design is to be used.

Table 5 --Comparison of area and multiple frame estimates, Idaho

Survey item	Area frame[1] CV %	Multiple frame CV %
Cattle inventory	7.9	
- Tract[2]	7.9	
- Farm[3]	9.4	
- Multiple frame		4.9 [4]
Potato acres		
- Tract	15.0	
- Multiple frame		5.8 [5]

1/ Area frame sample was 362 segments
2/ The tract or closed estimation is based on information physically located within the segment.
3/ The farm or open estimator uses data for the entire farm if the headquarters is located within the segment.
4/ List sample = 1,033 names
5/ List sample = 450 names

When multiple frame sampling is being considered, the costs of developing and maintaining a list frame need to be weighed against both of the above factors.

This also complicates the sample allocation to the two frames. For example, while supplementing the area frame with a list sample of cattle producers improves the sampling error, it may be that the area frame is just as efficient as the list frame for certain types or sizes of livestock operations. Since the area frame is developed and a basic sample must be screened, it is necessary to determine the optimum mix of the area and list frames. In some instances, the area frame may be efficient for small livestock producers. In that case, list development efforts can be directed to only maintaining a list of large operators. The determination of the allocation to the area and list frames is based on experience from many surveys. The basic procedure has been to further subdivide the area overlap domain into subdomains - each representing a stratum in the list frame. Then variance and cost considerations are used to determine the optimum cutoff for list frame development and sampling because the so called screening estimator is used. The estimator as developed by Hartley relies upon one weight for the entire area frame overlap domain. In practice this weight is very small.

A paper by Fuller and Burmeister (13) provides an excellent reference for most of the theoretical work on multiple frame estimation subsequent to Hartley's initial effort. Most of the recent work requires knowledge of the domain sizes. When an area frame is used, the domain sizes can only be estimated.

An extension to Hartley's estimator was provided by Bosecker and Ford (1). They showed that a multiple frame estimator with different weights for each subdomain in the area overlap domain results in smaller variances than a weight for the entire overlap domain. They showed that optimum weights attached to the area domains and list frame strata differ considerably between strata.

Considerable effort is still needed to determine the appropriate estimator and the allocation to sample frames.

POLICY ISSUES

Some policy issues relate to the construction and maintenance of sample frames.

First, SRS has built its estimating program upon the joint use of area and list frames for several reasons.

(1) The resources and capability to construct a complete list do not exist. SRS does not have access to many administrative lists. The Census Bureau does have access to administrative lists, but has been unable to construct a complete list. As reported by Dea et. al. (6), the Bureau assembled 19.0 million names from administrative sources from two different tax years and from the previous census. A duplication removal effort and a special farm identification survey were used to reduce the size of the file and to identify new operators. This effort resulted in a mail list of about 3.6 million names. In spite of this, they still missed about 15 percent of the farms as determined by a coverage evaluation survey.

(2) The estimates of production generated by the Department are based upon the premise that two measures are needed. One measure is that of the level or magnitude of production. The other measure is the measure of change over time. Both are equally important and one cannot be slighted for the other. The level of an estimate

or the measure of change should not be swayed by changes in the completeness of a list frame.

Another policy issue that crops up from time to time is "Who should maintain the farm list?" The same argument could also apply to the area frame. These arguments probably go beyond the scope of a technical paper and will not be further pursued.

REFERENCES

(1) Bosecker, Raymond R., and Barry L. Ford, "Multiple Frame Estimation with Stratified Overlap Domain." Proceeding of the Social Statistics Section, American Statistical Association, 1976.

(2) Ciancio, Nicholas J., Dwight A. Rockwell, and Robert D. Tortora. "An Empirical Study of Area Frame Stratification." U.S. Department of Agriculture, Statistical Reporting Service, July 1977.

(3) Clark, Cynthia Z.F. "Comparability of Data from Censuses of Agriculture." Proceedings of the Section on Survey Resarch Methods, American Statistical Association, August 1984.

(4) Cochran, William G. Sampling Techniques, 3rd Edition, New York: John Wiley and Sons, 1977.

(5) Craig M., R. Sigman, and M. Cardenas. "Area Estimates by LANDSAT: Kansas 1976 Winter Wheat." U.S. Department of Agriculture, Economics, Statistics and Cooperative Service, August 1978.

(6) Dea, Jane Y., Tommy W. Goulden, and D. Dean Proachaska. "Record Linkage for the 1982 Census of Agriculture Mail List Development Using Multiple Sources." Proceedings of the Section on Survey Resarch Methods, American Statistical Association, August 1984.

(7) Fecso, Ron. "Stratification of a Remotely Sensed Area Sampling Frame," Proceedings of the Survey Research Section, American Statistical Association, 1981.

(8) Fecso, Ron, Jeff Geuder, Bob Hale and Steve Pavlasek. "Estimating Dry Bean Acreage in Michigan." SRS Staff Report No. AGES820225. U.S. Department of Agriculture, Statistical Reporting Service, 1982.

(9) Fecso, Ron and Van Johnson, The New California Area Frame: A Statistical Study, SRS-22, U.S. Department of Agriculture, Statistical Reporting Service, 1981.

(10) Fecso, Ron, Van Johnson and Jeff Geuder. "Using SAS to Evalute an Area Sampling Frame for Agricultural Surveys." Proceedings of the Sixth Annual SAS Users Group International Conference. Orlando, Florida, Feb. 1981, 363-368.

(11) Fecso, Ron and Robert D. Tortora. "Area Frame Sampling in Agriculture: Developments and Prospects." Presented at the International Conference in Statistics: An Appraisal, Ames, Iowa, June 1983.

(12) Fellegi, J. P. and A. B. Sunter. "A Theory for Record Linkage," Journal of the American Statistical Association Vol 64, (1969). 1183-1210.

(13) Fuller, Wayne A. and Leon F. Burmeister. "Estimators for Samples Selected From Two Overlapping Frames." Proceedings of the Social Science Section, American Statistical Association, 1972.

(14) Geuder, Jeff. "Paper Stratification in SRS Area Sampling Frames." SF&SRB Staff Report No. 79. U.S. Department of Agriculture, Statistical Reporting Service, 1984.

(15) Gurney, Margaret and Maria Elena Gonzalez. "Estimates for Samples From Frames Where Some Units Have Multiple Listings," Proceedings, American Statistical Association, 1972.

(16) Hanuschak, George A., and Kathleen M. Morrissey. "Pilot Study of the Potential Contributions of LANDSAT Data in the Construction of Area Sampling Frames." U.S. Department of Agriculture, Statistical Reporting Service, Oct. 1977.

(17) Hartley, H. O. "Multiple Frame Surveys," Proceedings, American Statistical Association, 1962.

(18) Houseman, Earl. Area Frame Sampling in Agriculture, SRS-21. U.S. Department of Agriculture, Statistical Reporting Service, 1975.

(19) Huddleston, H.F., P.L. Claypool, and R.R. Hocking. "Optimal Allocation to Strata Using Convex Programming." The Journal of the Royal Statistical Society, Series C, Vol. 19, No. 3, (1970), 273-278.

(20) King, A.J. and R.J. Jessen. "The Master Sampling of Agriculture. Journal of the American Statistical Association Vol. 40, March 1945, 38-56.

(21) Pratt, William L. "The Use of Interpenetrating Sampling in Area Frames." U.S. Department of Agriculture, Statistical Reporting Service, May 1974.

(22) 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) pp. 87-90.

(23) U.S. Department of Agriculture, Statistical Reporting Service. Scope and Methods of the Statistical Reporting Service, Publication No. 1308, July 1975.