Journal of Official Statistics
Vol. 2. No. 3, 1986, pp. 279–292
© Statistics Sweden

Sampling Frames for Agriculture in the United States

Ron Fecso, Robert D. Tortora, and Frederic A. Vogel[1]

Abstract: The primary purpose of the National Agricultural Statistics Service (formerly the Statistical Reporting Service) in the U.S. Department of Agriculture is to provide information about current and future supplies of agricultural commodities. This information is generally obtained through surveys relying upon the joint use of area and list sampling frames. This paper provides an overview of the historic development and current use of the area and list sampling frames and their integration through the use of dual frame sampling techniques.

Key words: Agriculture; area frame; list frame; dual frame.

1. Introduction

Farmers, ranchers, and agribusiness firms need information about current and future supplies of agricultural commodities for marketing, planning, and decisionmaking. This information is also necessary for policy decisions concerning Government programs affecting the agricultural economy in specific ways and the United States and global economies in more general ways. To meet these needs the National Agricultural Statistics Service (NASS) (formerly the Statistical Reporting Service) of the U.S. Department of Agriculture (USDA) conducts sample surveys and publishes about 300 national and 9000 State reports each year. These reports cover a broad range of agriculture crops, livestock, and economic items (USDA (1983)).

Agriculture in the United States is a business that consists of about 2.2 million farms and ranches. These operations vary in size and in the types of commodities they produce. For example, farms and ranches vary widely in size as measured by total value of production. One-third of the operations account for over 90 percent of the total value of production. (A farm or ranch is a place producing $1000 or more per year of agricultural products.) One percent of the operations account for a third of the total sales. But, the operations differ considerably in what they produce. Only 10 percent of the farms account for three-fourths of the corn acres. Less than three percent produce crops such as peanuts, cotton, or rice.

Thus, farms and ranches in the United States present two sampling problems for NASS. First, the population is highly skewed, with a large number of small operations and a few very large operations. Second, many of

[1] Head, Yield Assessment Section; Director, Survey Division; Director, Statistical Research Division, respectively, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250, USA. The views expressed herein are not necessarily those of NASS or USDA. The authors would like to thank members of the Statistical Research Division, the editor, and referees for their many valuable comments and suggestions.

the commodities that the farms and ranches produce are "rare items." In this paper, a rare item is any agricultural commodity that is produced on only a small proportion of the operations in a State. Because of the highly skewed population and rare agricultural commodities, NASS uses dual frame (area and list) sampling procedures.

The area frame is complete in the sense that all farms have a known positive probability of selection. The frame is suitable for general-purpose surveys that cover a wide spectrum of crop and livestock items. It can also be used for economic-type surveys where defining the reporting unit for a farm is more difficult. Although the initial investment in developing the frame is large, the lifespan of an area frame is long, usually more than 15 years.

A weakness of the area sampling frame is that it is inefficient for commodities on large farms or commodities that are rare. Some items such as cattle, which in many areas are produced on a large proportion of the farms, are also characterized by large variability in 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 having a distribution which is less variable than most commodities. However, only 0.5 percent of the farms produce rice. This means that a general-purpose sample of area frame segments would yield only about 1 in 200 farms actually reporting rice unless some crop-specific stratification and disproportionate allocation were employed.

Because area frame sampling has limitations, NASS 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 start. In 1882, part-time statisticians were appointed to develop and maintain groups of voluntary crop reporters to provide current information about agriculture. In 1892, 125000 farm operators were furnishing survey data for annual estimates. The generalized structure of agriculture through the middle of the 20th century allowed the United States 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.

As agriculture became more specialized and large operations began emerging, NASS began shifting its survey program away from the general-purpose nonprobability surveys to area frame probability surveys and a search for procedures to supplement the area frame. Research by Hartley (1962) led to the implementation of dual frame sampling using the area and list frames.

For NASS, the two most important characteristics for list frames are:

Coverage–A large portion of the agricultural item(s) being estimated should be covered for efficient allocation.

Measure of size–Most of the sampling units should have a relative measure of size for the item(s) being estimated for efficient stratification.

As in most establishment surveys, the task of compiling a complete agricultural list is impossible. Therefore, the main strength of a list frame is to supplement the area frame's weaknesses – estimation of rare items and items with extreme variability. This paper discusses some of these strengths, provides an overview of the area and list frames, and outlines current and proposed research at NASS on area and list frames.

2. Area Frame Overview

We now discuss two aspects of area frame sampling in agriculture:

1) developments in the construction and sample design since development of the master sample of agriculture (King and Jessen (1945)), and
2) prospects for improving the construction and maintenance of area sampling frames (Fecso and Tortora (1983)).

Since 1967, NASS 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 specialty surveys (USDA (1983)). The current area frame evolved from the master sample of agriculture. The master sample frames were constructed on county highway maps with minor civil divisions and sample units delineated on these maps. On the average, each sample unit contained about four farms. Crop reporting districts within each State were used to provide geographic stratification. Changes in the area frame design were adopted slowly over the 35-year period from 1940 to 1975. The changes included a refined stratification process and the introduction of replicated sampling. Until the mid-1970's, the area frame construction and maintenance process could be characterized as being the same paper-and-pencil operation, using the same types of materials as used for the master sample of agriculture. After 1975, however, new technologies affected the area frame construction process. NASS started using the computer at several points in the process, from measuring the land area of the frame and selecting the sample to providing quality control for frame construction (Fecso and Johnson (1981) and Fecso, Johnson, and Geuder (1981)). Table 1 summarizes the significant chronological events in area frame sampling for agriculture. Notice that the changes made in the 1960's and early 1970's are primarily related to sampling methods, such as the new stratification by land use and the introduction of replicated sampling. The changes that began in the mid-1970's, on the other hand, represent the application of new technology to area frame construction, such as the use of the computer and the availability of satellite imagery.

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 National Agricultural Statistics Service begins investigating the use of area frame sampling
1962	Land-use stratification is introduced in NASS area frames
1967	All 48 conterminous States have area frames
1973	Replicated 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

Table 1 (cont.).

	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
1986	Geographic Information System Development begins

Table 2. The Master Sample and Land Use Area Frames (1945 compared to 1986)

Frame characteristic	Master sample	Land use area frame
Year of frame	1945	1986
US number of farms	6 million	2.2 million
Sample size	67000 segments	16000 segments
Resident farm operators in sample	300000	16000
Measure of size	Indicated number of farms	Area of sample unit (segment)
Stratification	By Crop Reporting District: Urban Places Rural Places Open Country	Land Use Potential Urban Crop Specific

In 1978, NASS replaced the last master sample frame with frames stratified according to land use. Some of the characteristics of the master sample frame and the current NASS area frames are given in Table 2 for comparison purposes.

There was about a 60 percent drop in the total number of U.S. farms and ranches between 1945 and 1986. The new frames differ in that the number of sample segments has decreased by over 75 percent. This corresponds with a decrease of about 95 percent in resident farm operators.

NASS builds area frames on mosaics of aerial photographs and then transfers boundaries to county highway maps to measure land areas accurately. Table 3 lists the steps in area frame construction and shows where NASS uses the computer in this process. In general, the Urban and Rural Places strata of the master sample are still being used by NASS in its land-use frames. However, the Open Country stratum has been further subdivided to obtain improved sampling efficiency. To begin land-use stratification in a State, blocks of similar areas of land are identified within each county (counties and political subdivisions are used as a tool to manage the work flow of area frame construction) 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, 5) nonagricultural land such as parks and military reservations, and 6) water. 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.

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 (psū's), on aerial photographs and county highway maps
Digitization	The area of the psu is measured through use of a microcomputer and digitizing tablet
County Level Edits	Data is transferred to a minicomputer for 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 probability proportional to size 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 sampling units (segments) on aerial photography

Table 4. Strata 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, extensively cultivated land in Canyon, Ada, Owyhee, Elmore, Gooding, Twin Falls, Lincoln, Jerome, Mindoka, Cassia, Power, Bannock, Caribou, Bingham, Bonneville, Teton, Madison, Jefferson, Fremont, Clark, and Butte 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 area that is not in dryland grains.
22	Dryland Grains – 15 to 33 percent cultivated. Extensively cultivated land used in conjunction with stratum 10. (May be 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.

After stratification the primary sampling units (psu's) are defined on the photo mosaics. The average size of a psu varies by stratum. For agricultural strata, a psu contains an average of eight sampling units or segments, each of which is approximately one square mile in size. During the construction of the primary sampling units, the main emphasis is on delineating units that can be further subdivided into segments with observable boundaries. Enumerators need to find these boundaries easily during data collection. Clear boundaries are important to minimize nonsampling errors, but they can conflict with the desire to have a homogeneous distribution among the segments to minimize sampling variation. Each primary sampling unit is digitized; that is, the unit's outline on a county highway map is stored on a computer as a series of boundary coordinates. A computer program is then used to obtain the area of each psu and to plot each county to ensure that each psu has been digitized and assigned to its proper stratum. After a State is completely digitized, another computer program selects a sample of first-stage units. Each psu selected in the sample is further subdivided on the photo mosaic. One of the subdivided areas is selected at random from each selected psu. The point to be noted here is that a two-step selection procedure is used which reduces the costs of frame building. Except for unusally large segments in rangeland areas and for some segments in 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 was used. The last section outlines the prospects for these processes in the future.

3. Area Frame Developments

Area frame construction is a major undertaking which must be considered a long-term investment. The 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, NASS maintains an ongoing research effort to improve area frame sampling. The research in area frame sampling has been directed toward the search for cost-saving techniques and methods to improve the efficiency of the estimators. This section will outline the major changes to the NASS area frame made since the master sample was used.

3.1. Stratification

One of the first major changes to the master sample concepts was stratification by land use. Starting in the early 1960s, 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 dense residential development (Ciancio, Rockwell, and Tortora (1977)).

By 1978, all States had area frames with a form of land-use stratification. These frames continue to be updated at the rate of two or three States 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 which are a source of nonsampling errors, and the land use within strata changes. Because the NASS area frame is used to collect multiple data items, there has been much debate over the most effective use of stratification (Ciancio, Rockwell, and Tortora (1977), Fecso, Geuder, Hale, and Pavlasek (1982), Fecso and Johnson (1981), Hanuschak and Morrissey (1977), Houseman (1975), and Huddleston, Claypool, and Hocking (1970)). Stratification for more than a few specific commodities is difficult. A sample allocation which is optimal for one commodity could reduce the efficiency for other commodities. 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's new frame (Fecso, Johnson, and Geuder (1981)). As a result of this effort NASS is using crop-specific stratification in States with concentrations of important crops when those crop-specific areas can be identified on available materials. Examples of crop-specific strata include fruits and vegetables in California (new frame in 1979), dryland grains in Washington, Oregon (1980), and Idaho (1982), and rice, cotton, wheat, and peanuts in Texas (1982).

Crop-specific stratification can improve multicommodity survey designs (Fecso and Johnson (1981)). Creating certain crop-specific strata results in more efficient estimation of the specified crop even with a lower sampling rate in the crop-specific strata. Sampling efficiency improves for other crops since the sampling rate can be increased in the remaining general strata. California is an example of the gains observed using these techniques. The relative efficiency of the acreage estimates for major commodities was about the same in the new frame as the old frame, but the area of the sample which was enumerated was about double in the old frame. Thus, considerable reductions in cost for fixed variance are possible.

3.2. Replicated sampling

For all frames constructed since 1974 NASS uses replicated sample designs (Pratt (1974)). Replicated sampling is characterized by the selection of several independent samples from the frame. It was started to facilitate the rotation of sampling units to limit 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 developing nations and for special surveys).

Replicated sampling, as done by NASS, uses a form of substratification called "paper stratification" – essentially a geographic substratification of each State (Geuder (1984)). The first step in paper stratification is to determine an ordering of the psu's in each stratum. To determine this ordering, NASS groups counties into "similar" agricultural areas using cluster analysis. Similar counties are put in sequence by 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 divisible by the number of paper strata). Strata with few sampling units usually have two or three paper strata, while large strata may have from 10 to 20 paper strata.

A replicate in the NASS design is defined as a simple random sample of one segment from each paper stratum in a land-use stratum. Thus, the paper strata serve much the same purpose as a systematic sample, dispersing the sample throughout the population, but they also are a form of commodity-specific stratification which increases the efficiency of the estimates.

3.3. Materials

An important problem in NASS 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 decrease the efficiency of the land-use stratification and create problems for enumerators because the old photography does not indicate the current land use. Eventually, enough gain in efficiency can be realized from restratification to justify the cost of a new frame. To get the most gain from a new frame, current materials are essential. Prior to 1979, NASS stratified frames using photo mosaics. These materials were often 20 years old in some areas and rarely less than three years old. Also, in some areas, no aerial coverage was available and topographic maps had to be used. To overcome the problems of old or missing photo coverage, NASS started using Landsat satellite imagery in 1979 to aid in stratification. But Landsat imagery does not provide enough detail to create the frame; it is used to update old materials to reflect current land use.

Easily identified segment boundaries are a requirement for effective enumeration. Enumerator identification of boundaries improved with a new program for high-altitude color infrared photo coverage, the National High Altitude Photography (NHAP) program, a cooperative venture by various government agencies. The NHAP project, started in 1980, provides complete coverage of the continental United States every 5 or 6 years. NASS now uses this material for frame construction and for data collection.

3.4. Sample allocation

NASS continues research on methods to allocate the most efficient samples to new frames. 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 average interview times for each stratum so that cost data is incorporated in the allocation formulas (Fecso and Johnson (1981)).

3.5. Quality control

Quality control of the NASS area frame is achieved in several ways and the development of improved controls continues. Prior to 1978, most quality control was a result of a post-survey review of the variances, the photographic materials used by the enumerator to find segments, and the completed questionnaires. In 1978, NASS started using quality control procedures to ensure that segments are randomly selected and that frames conform to specifications. In 1979, a post-survey analysis package was developed which helped point out statistically inefficient frame construction techniques as well as construction errors (Fecso, Johnson, and Geuder (1981)). By 1980 a minicomputer was being used to improve the quality of the sample frame by finding problems prior to the sample selection rather than in post-survey analysis.

3.6. Another area frame design

From 1979 to 1982 NASS built area sampling frames to estimate corn and soybean acres in portions of Argentina and Brazil (Fecso (1981)). These frames were built using satellite imagery and navigational maps. The frames were developed, under a joint research agreement with the National Aeronautics and Space Administration, to attempt to estimate corn and soybean acreages without any ground data collection. After the frames were developed, estimates would be made using image interpretation and digital classification of Landsat scenes. Although budgetary priorities changed before estimates were made, NASS developed methods for and gained considerable experience in using remote sensing and computer technology for frame construction. Because of this research, NASS now uses remote sensing and computer technology for domestic frame construction. One example is NASS's use of Landsat imagery to improve stratification for estimating specialty crops. Another example is the use of a system of microcomputers, digitizers, and a minicomputer for quality control and automation of frame construction.

3.7. *New frame analysis*

Before constructing a new frame, NASS assembles and analyzes a considerable amount of information to help increase precision 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 estimates on national precision, reviewing county estimates, gathering data on urban development and changes in land usage, and analyzing prior years' data. The information is used to determine the type of stratification which would be most efficient in the new frame. After stratification, the prior years' segments are located on the new frame and post-stratified to provide an estimate of the stratum variances for 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 NASS completes the first survey using the new frame, the data is processed through the Area Frame Analysis package (Fecso, Johnson, and Geuder (1981)). This package provides graphical and statistical information for an analysis of the frame construction and the sources of variation. The analysis uncovers nonsampling errors, improves allocations, and gives insight into future design or construction alternatives.

4. Area Frame Prospects

NASS increased resources for area frame research in 1978. Several projects outlined in this section will have short- and long-range impacts on area frame construction and use.

4.1. *Landsat*

Landsat imagery and its digital data have the greatest potential for improving area frame sampling (Craig, Sigman, and Cardenas (1978), Fecso and Johnson (1981), and Hanuschak and Morrissey (1977)). As processing costs decline and classification of multispectral data into land cover classes improves, several major changes could take place. The most immediate is the use of Landsat technology in second-stage sampling. Here a psu is divided into a predetermined number of sampling units, retaining observable boundaries while making units of nearly equal size. Landsat data for the psu could be used to achieve a homogeneous divison of units. This process would reduce the variance of crop acreage estimates from area frame surveys.

Research using Landsat data to detect land use changes has several possible applications. Auxiliary data for the primary and secondary sample units can be used in regression estimators for crop acreage. These estimators should be more precise than the direct expansion estimates, if NASS can acquire the Landsat data in a timely and cost-effective way. Another approach is to use Landsat data for stratification. Timeliness, efficiency, and cost may all be improved by using the Landsat data as auxiliary data for psu's to improve stratification and sample allocation. Current Landsat data could be used for post-stratification, providing NASS with the ability to increase the efficiency of surveys as agricultural practices change.

4.2. Geographic information system

When psus are digitized, the area frame becomes a Geographic Information System (GIS). With auxiliary digital Landsat information, the area GIS can be maintained like a list frame. The registration of psu boundaries to a geographic reference system, such as latitude-longitude coordinates, can allow the automatic updating of each psu as land-use changes, e.g., farmland to housing, rangeland to irrigated cropland, etc. This new information can be used for restratification. As the psus are updated, the impact of their changes can be evaluated. When there is sufficient change in the auxiliary data to show that sample design changes would improve the efficiency of the estimators, the restratification process can be started.

When an area frame becomes a GIS, a significant portion of the construction process can be moved from the manual to the automated mode.Referring to Table 3, we see that most of the process up to segment selection could be done by machine.

4.3. Frame rotation

The NASS area frame survey design calls for a 20 percent replacement of segments each year. When NASS builds new area frames, many segments are not used for the full rotation cycle: those segments in the old frame which were not in the sample for the full five-year rotation cycle and those in the new frame which will be rotated out during the first few years. For example, in each year from 1979–83, about 1050 old frame segments out of about 16000 sampled nationally were rotated out before the full cycle, while an average of about 975 new frame segments took their place. Since 80 percent of each group does not receive full use, the concept of rotating into a new frame has the potential for cost savings as well as minimizing possible level changes in estimates from the new frame. "Rotating into a new frame" means that NASS would need to use a combined estimator, using a weighted average of independent estimates from each frame.

The weights might be chosen as the proportion of the total segments sampled in each frame. Thus, the weight for the old frame decreases as the rotation scheme removes 20 percent of the old frame segments and draws an additional 20 percent of the segments from the new frame. Other choices also exist. Considering that the new frame should be more efficient, the weight for the old frame might be chosen smaller.

Frame updates would also be facilitated by rotating into a new frame. Whenever new auxiliary information is obtained, an automated update can be started. If frame errors are found or if the first-year allocation is inefficient, redesign is done for the small, new sample rather than a full new-frame sample.

This section shows that future NASS area frames will be built using staff from many disciplines: cartography, geography, data processing, remote sensing and statistics. These new frames will provide more efficient estimates and have reduced nonsampling errors.

5. List Frame

For agriculture, a list frame is either a list of producers or agricultural businesses. The list frame should contain names, addresses, telephone numbers, and measures of size. Ideally, the list should be complete and free of duplication. Neither condition is satisfied in practice. 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.

Since the process of constructing a list frame involves assembling lists from a variety of sources, identifying duplication is a major problem. The matching 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, addresses, and other information. The underlying theory for the model used by NASS was developed by Fellegi and Sunter (1969). By using statistical record linkage, two probability values (threshold values) are used to assign records to one of three groups:

Nonlinked records
Probable linked records
Definite linked records

The threshold values are used to separate nonlinked records from definite links. All probable linked records are reviewed before sampling takes place. Manual resolution is, however, 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.

There are several alternatives for handling the duplication that remains in the frame and is not detected until the sample has been selected. One solution is presented by Gurney and Gonzales (1972) for the case where the number of times a given operation is duplicated is known. Another method has been developed by Rao (1968) for the case when the number of times an operation can be selected from the frame is unknown.

NASS attempts 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 detect possible duplication. Each respondent is asked whether the farm or ranch is known by any operation name or if any other names (partners) are associated with the operation.

The other problem of a list frame is incompleteness. Not only do the contents of the frame change, names enter and exit agriculture. But the operations change in their structure and size from year to year. About 20 percent of the records in the U.S. agricultural list frame will change from year to year. Therefore, a savings from collecting data from list samples should exceed frame maintenance costs.

NASS uses the area frame to measure the incompleteness of the list frame in several ways, including:

number of farms and ranches,
land in farms and ranches,
value of agricultural sales,
number of cattle and calves,
number of hogs and pigs, and
total grain in storage.

All of these different measures of incompleteness are useful to NASS. For example, Hill and Steiner (1985) found that while in 1984 the NASS list frame included about 54 percent of the farms, it accounted for over 75 percent of the land in farms, total cattle and calves, and total hogs and pigs. Thus, the NASS list frames have a higher proportion of the large farms and ranches.

6. Dual Frames

The primary reason for using dual or multiple frame sampling procedures is to capture the strengths of the area and list frames. The list frame, while incomplete, can be sampled efficiently. 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. NASS uses the area frame with list frames to estimate a variety of commodities. A comprehensive reference for dual frame procedures used by NASS is Nealon (1984).

Dual frame surveys are subject to all operational problems that occur with single frame surveys. See (Vogel (1975)). By their very design, problems unique to dual or multiple frame surveys also occur.

The overlap between frames must be determined. NASS must decide whether every area frame reporting unit is on the list frame. 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 dual frame estimation:

(1) *Nonoverlap Domain* – This domain consists of producers in the area frame sample that are not in the list frame.

(2) *Overlap Domain* – This domain consists of producers in the area frame sample 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 area frame estimate of characteristic X can be written as:

$$X'_{\text{area}} = X'_{\text{nol}} + X'_{\text{ol}},$$

where X'_{nol} is an estimate of the total for the portion of the population missing from the list frame (the nonoverlap domain of the area frame) and X'_{ol} is the area frame estimate of the population also represented by the list frame (overlap domain).

A dual frame estimator of X proposed by Hartley (1962) is

$$X' = X'_{\text{nol}} + PX'_{\text{ol}} + QX'_{\text{l}},$$

where 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 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 a major source of nonsampling errors in NASS surveys. The name-matching operation can be completed manually or automatically using decision logic about what is a match and what is a nonmatch.

The sampling efficiencies to be gained through dual frame sampling are illustrated with data from the State of Idaho. The gains from dual frame estimation with a list sample of 450 names of potato producers are large. The area frame sample of 362 segments, which had a coefficient of variation (CV) of 15 percent for estimated potato acres, would need an increase in the number of segments by a factor of nine to achieve the same sampling precision provided by the dual frame estimate (5.8 percent). The usefulness of dual frame sampling for specialty crops is clear if the size of the area frame sample is adequate for other items being estimated. A comparison of the cattle inventory estimates, which had a list sample of 1033 names, found an area frame CV of 8 percent and a multiple frame CV of 5 percent.

Several factors need to be evaluated when considering the use of dual frame sampling. For example, with agricultural surveys in the United States 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:

Those items only estimated for by the area frame and

Those items for which the area frame estimate is not adequate – where the incompleteness of the list frame is measured from the area frame.

When dual frame sampling is being considered, the costs of developing and maintaining a list frame need to be weighed against the above factors. For example, while supplementing the area frame with a list sample of cattle producers reduces the sampling error, the area frame may be as efficient as the list frame for certain types or sizes of livestock operations, such as small livestock producers. In that case, list development efforts can concentrate on maintaining a list of large operators.

The paper by Fuller and Burmeister (1972) is an excellent reference for most of the theoretical work on dual frame estimation after 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 (1976). They showed that a dual 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 by strata. But more work is needed to improve dual frame estimators and allocations to sample frames.

7. Summary

This paper discusses the sampling frames used by NASS of the U.S. Department of Agriculture. The paper give a historic overview of the frame development for U.S. agriculture, describes the current methods NASS uses in area and list frame construction, maintenance and sampling, and describes future developments for sampling frames in the United States.

8. References

Bosecker, R. R. and Ford, B. L. (1976): Multiple Frame Estimation with Stratified Overlap Domain. American Statistical Association, Proceedings of the Social Statistics Section, pp. 219–224.

Ciancio, N. J., Rockwell, D. A., and Tortora, R. D. (1977): An Empirical Study of Area Frame Stratification. U.S. Department of Agriculture, Statistical Reporting Service Staff Report. Washington, D.C.

Craig, M., Sigman, R., and Cardenas, M. (1978): Area Estimates by LANDSAT: Kansas 1976 Winter Wheat. U.S. Department of Agriculture, Economics, Statistics and Cooperative Service Staff Report. Washington, D.C.

Fecso, R. (1981): Stratification of a Remotely Sensed Area Sampling Frame. American Statistical Association, Proceedings of the Survey Research Section, pp. 506–508.

Fecso, R., Geuder, J., Hale, B., and Pavlasek, S. (1982): Estimating Dry Bean Acreage in Michigan. U.S. Department of Agriculture, Statistical Reporting Service Staff Report No. AGES820225. Washington, D.C.

Fecso, R. and Johnson, V. (1981): The New California Area Frame: A Statistical Study. U.S. Department of Agriculture, Statistical Reporting Service Publication SRS-22. Washington, D.C.

Fecso, R., Johnson, V., and Geuder, J. (1981): Using SAS to Evaluate an Area Sampling Frame for Agricultural Surveys. Proceedings of the Sixth Annual SAS Users Group International Conference, Orlando, Florida, USA.

Fecso, R. and Tortora, R.D. (1983): Area Frame Sampling in Agriculture: Developments and Prospects. Presented at the International Conference in Statistics: An Appraisal, Iowa State University, Ames, Iowa, USA.

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

Fuller, W. A., and Burmeister, L. F. (1972): Estimators for Samples Selected From Two Overlapping Frames. American Statistical Association, Proceedings of the Social Statistics Section, pp. 245–249.

Geuder, J. (1984): Paper Stratification in SRS Area Sampling Frames. U.S. Department of Agriculture, Statistical Reporting Service, SF&SRB Staff Report No. 79, Washington , D.C.

Gurney, M. and Gonzalez, M. E. (1972): Estimates for Samples From Frames Where Some Units Have Multiple Listings. American Statistical Association, Proceedings of the Social Statistics Section, pp. 283–288.

Hanuschak, G. A and Morrissey, K. M. (1977): Pilot Study of the Potential Contributions of LANDSAT Data in the Construction of Area Sampling Frames. U.S. Department of Agriculture, Statistical Reporting Service Staff Report, Washington, D.C.

Hartley, H. O. (1962): Multiple Frame Surveys. American Statistical Association, Proceedings of the Social Statistics Section, pp. 203–206.

Hill, G.W. and Steiner, M. A. (1985): List Frame Evaluation From an Area Frame Perspective. U.S. Department of Agriculture, Statistical Reporting Service Staff Report, Washington, D.C.

Houseman, E. (1975): Area Frame Sampling in Agriculture. U.S Department of Agriculture, Statistical Reporting Service Publication SRS-21, Washington, D.C.

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

King, A. J. and Jessen, R. J. (1945): The Master Sampling of Agriculture. Journal of the American Statistical Association, Vol. 40, pp. 38–56.

Nealon, J. P. (1984): Review of the Multiple and Area Frame Estimators. U.S Department of Agriculture, Statistical Reporting Service, SF&SRB Staff Report No. 80, Washington, D.C.

Pratt, W. L. (1974): The Use of Interpenetrating Sampling in Area Frames. U.S. Department of Agriculture, Statistical Reporting Service Staff Report, Washington, D.C.

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

U.S. Department of Agriculture (USDA) (1983): Scope and Methods of the Statistical Reporting Service. Publication No. 1308. Washington, D.C.

Vogel, F. A. (1975): Surveys with Overlapping Frames – Problems In Application. American Statistical Association, Proceedings of Social Statistics Section, Washington, D.C., pp. 694–699.

Received January 1986
Revised July 1986