THE AREA FRAME: A SAMPLING BASE FOR ESTABLISHMENT SURVEYS

Jeffrey Bush, Carol House, National Agricultural Statistics Service
Jeffrey Bush, NASS, Room 301, 3251 Old Lee Highway, Fairfax, Virginia 22030

KEY WORDS: Area frame, stratification, primary sampling unit

Area frame methodology has formed one of the cornerstones of probability sampling for several decades. While area frames are frequently used in urban settings for household surveys and population censuses, those with a rural focus have proved valuable for targeting farm establishments to provide basic statistics on agriculture and ecological resources. Cotter and Nealon outline the advantages and disadvantages of area frame methodology. They state that area frames are highly versatile sampling frames providing statistically sound estimates based on complete coverage of land area. Although costly to build they are generally slow to become outdated. However, area frame sampling is generally less efficient than list sampling for targeting any individual item and is inadequate for estimating rare populations.

This paper describes four different area frame methodologies currently in use as a base for sampling in rural areas. These are: a) the area frame used by the United States Department of Agriculture's (USDA) National Agricultural Statistics Service; b) the area frame used by Statistics Canada for agricultural statistics; c) the hexagonal area frame used by the U.S. Environmental Protection Agency; and d) the area frame used by the USDA's Soil Conservation Service for the National Inventory Survey. The paper's greatest emphasis is on the NASS area frame. For this frame, the authors provide additional detail on area frame construction, sampling, data collection and estimation. The paper provides a profile of costs associated with these activities, as well as procedures to assess quality deterioration in an "aging" frame.

NASS AREA FRAME

The National Agricultural Statistics Service (NASS) is the major data collector for the U.S. Department of Agriculture. As such it has responsibility to provide timely and accurate estimates of crop acreages, livestock inventories, farm expenditures, farm labor and similar agricultural items. NASS also provides statistical and data collection services to other Federal and State agencies. They have used an area sampling frame extensively for over 30 years in the pursuit of these objectives. Area frame samples are used alone and in combination with list samples (multiple frame). NASS contacts approximately 50,000 farm establishments each year through their area frame sampling procedures.

This section updates the work of Cotter and Nealon as it describes the procedures used by NASS to construct area frames and sample from them. It discusses data collection procedures, estimators, and costs associated with these different activities. Finally it discusses methods to objectively assess the "aging" of an area frame.

Area Frame Construction and Sampling

NASS constructs area frames separately by state and maintains one for every state except Alaska. Generally, two new frames are constructed each year to replace out-dated ones. The most recent frame construction was for Oklahoma. It became operational in June 1993. This frame will be used as the "example" throughout this paper.

Frame construction produces a complete listing of parcels of land, averaging six to eight square miles in size, throughout a state. These parcels serve as the primary sampling units (PSUs) in a two stage design, and each contain a varying number of population units or segments. The sampling process selects PSUs, and only those selected PSUs are broken down into segments. This two stage process saves considerable time and money over that required to break the entire land area into segments.

Construction of and sampling from an area frame involves five basic steps: 1) determining specifications for the frame; 2) stratifying the land area and delineating PSUs within each stratum, 3) allocating stratum level optimal sample sizes; 4) creating sub-strata and selecting PSUs; and 5) selecting segments within PSUs. Each step is discussed in detail below.

Frame Specifications

The specifications for building an area frame consist of strata definitions and target sizes for both PSUs and segments within each stratum. Statisticians define these by examining previous survey data, and assessing

urbanization and other trends in the state's agriculture. Table 1 lists the frame specifications for the Oklahoma frame.

Strata are based on general land usage. A typical NASS area frame employs one or more strata for land in intensive agricultural (50 percent or more cultivated), extensive agricultural (15 to 50 percent cultivated), and range land (less than 15 percent cultivated). Less frequently an area frame contains "crop specific" strata. This occurs when a high percentage of the land in a state is dedicated to the production of a specific type of crop, such as citrus in Florida. In addition, each area frame uses an agri-urban and commercial stratum (more than 100 homes per square mile) plus a non-agricultural stratum including such entities as military bases, airports, and wildlife reserves. Finally, large bodies of water are separated into a water stratum.

Boundary points for agricultural strata are generally restricted to a set of standardized breaks: 15, 25, 50 and 75 percent cultivated. To determine the exact breaks for a given state, the percent cultivated for each segment sampled under the old frame is calculated from survey data. The resulting distribution is examined using the cumulative square root of frequency rule proposed by Dalenius and Hodges. The standardized breaks may be collapsed or expanded based on the structure of the distribution.

Two criteria are the most important for determining target sizes for PSUs and segments within strata: availability of good natural boundaries and the expected number of farm establishments. Generally, a lack of good boundaries will prompt the use of larger target sizes, while a large expected number of farm establishments will prompt smaller target sizes.

Stratification and Delineation of PSUs

Once strata definitions are set, the stratification process divides the land area of the state into PSUs and assigns each to the appropriate land use strata. Each PSU must conform to the definition and target size outlined for its particular stratum. PSU boundaries become a permanent part of the area frame and must be identifiable for the life of the frame. Thus the stratifier uses only the most permanent boundaries available when drawing off PSUs. Acceptable boundaries include permanent roads, rivers, and railroads. The final product of the stratification process is a "frame" file which contains a record for every PSU in a state. Specifically, each record includes the PSU number, stratum assignment, county, and size. This frame file is maintained over the life of a frame as the sampling base.

Prior to 1990 the process of stratification used paper maps, aerial photography, satellite imagery, and a considerable quantity of skilled labor. The end product was a frame delineated on paper U.S. Geological Survey 1:100,000 scale maps. In 1990, NASS implemented its Computer Aided Stratification and Sampling System (CASS). CASS automates the stratification steps on a graphical workstation using digital satellite imagery and line graph (road and waterway) data from the U.S. Geological Survey.

The digital satellite imagery employed by the CASS system is currently obtained from the thematic mapper (TM) sensor on the LANDSAT-5 satellite. The TM has a spatial resolution of 30 meters and is made up of 7 spectral bands. TM bands 1-5 and 7 reside in the reflective region of the spectrum while band 6 is located in the thermal infrared region. NASS experience with the imagery shows that bands 2, 3, and 4 highlight cultivated areas of land most accurately.

Table 1: Oklahoma Frame Specifications

Stratum	Definition	Primary Sampling Unit Size			Segment Size
		Minimum	Desired	Maximum	
		(sq. miles)	(sq. miles)	(sq. miles)	(sq. miles)
11	>75% CULTIVATED	1.00	6.0 - 8.0	12.0	1.00
12	51-75% CULTIVATED	1.00	6.0 - 8.0	12.0	1.00
20	15-50% CULTIVATED	1.00	6.0 - 8.0	12.0	1.00
31	AGRI-URBAN:>100 HOME/SQMI	0.25	1.0 - 2.0	3.0	0.25
32	COMMERCIAL:>100 HOME/SQMI	0.10	0.5 - 1.0	1.0	0.10
40	<15% CULTIVATED	3.00	18.0 - 24.0	36.0	3.00
50	NON-AGRICULTURAL	1.00	none	50.0	pps[i]
62	WATER	1.00	none	none	not sampled

i. Segments are selected with probability proportional to size; i.e. PSU's are treated as segments.

While TM data is very useful in providing information with respect to land usage, its large scale (30 meter resolution) renders it practically useless for identifying good PSU boundaries. Therefore the CASS system also uses digital files of U.S. Geological Survey 1:100,000 scale maps, in which feature class codes are assigned to all roads, water, railroads, power lines, and pipelines. The CASS system incorporates the road and waterway data from these files and overlays it on the TM imagery.

Personnel use a mouse and a "drawing" program to delineate boundaries of the PSUs and label them with their appropriate stratum number and sequence. As each PSU is completed, its size is immediately displayed. If the PSU does not fall within the particular target size, the stratifier immediately makes a correction. In addition, once a county has been completely divided into PSUs, the system can check for overlaps or omissions of land. Though the software provides many quality checks which save much time, reviews are still necessary to check the quality of stratification.

Figure 1 displays the stratification of Muskogee, OK which was performed on the CASS system. Notice the differing PSU sizes created with respect to each stratum.

Sample Allocation, Sub-stratification and Sample Selection

The national sample size for the NASS frame is approximately 15,000. The two stage design selects 15,000 PSUs and then one segment per selected PSU. The sampling process is described in more detail below.

Following stratification a multivariate optimal allocation analysis is performed to allocate the first stage sample of PSUs between land use strata. This is a multivariate procedure because the area frame must target a variety of agricultural items. The analysis requires the following inputs: a) population counts of segments per stratum; b) estimated totals of important commodities from previous year's survey; c) standard deviations from previous survey derived by locating old segments in the new frame and strata; and d) target CV's for major commodities. The analysis produces stratum level sample sizes with expected coefficients of variation less than or equal to the target CV's.

At this point in the process, PSUs have been delineated and stratified according to their land use and the optimal number of PSUs to sample in each stratum has been determined. Next, each land use strata is further divided into sub-strata based, in part, on a criteria of agricultural similarity. This process improves the precision of estimates of individual commodities and facilitates sampling by replication. PSUs are grouped by county, and ordered within counties in a serpentine pattern starting in the North East corner. Counties are then ordered based on results of a clustering algorithm that groups counties with similar crop production. Together these steps produce an ordering of all PSUs throughout the state. Figure 2 displays the county ordering for the

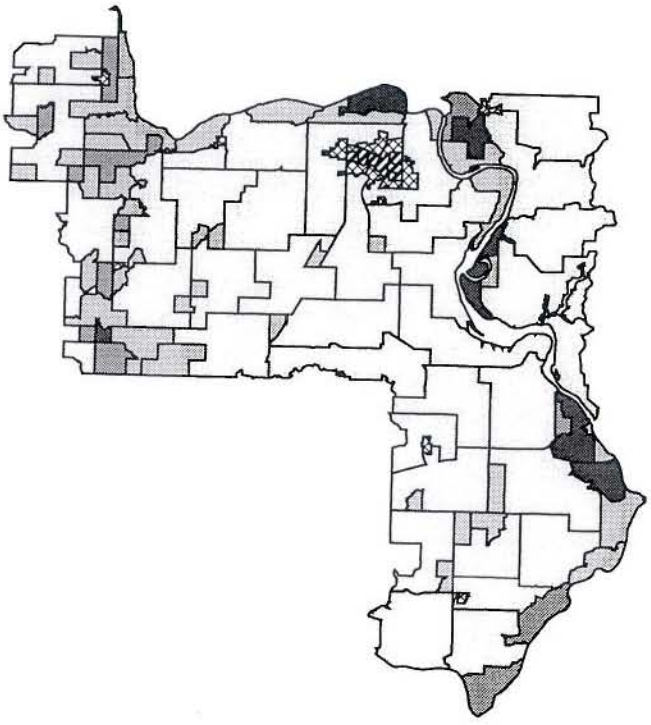

Figure 1: Stratification of Muskogee County, OK

state of Oklahoma. Substrata then equally divide the ordered PSUs within each stratum, where one PSU is selected per replicate per sub-stratum.

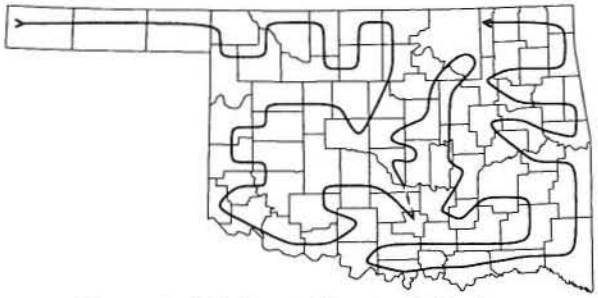

Figure 2: Oklahoma County Ordering

The sampled PSUs in each sub-stratum are randomly selected with probability proportional to the number of segments they contain and are assigned to a replicate. In non-agricultural and some range strata, where the lack of suitable boundaries is a problem, the PSUs themselves also serve as segments.

Replicated sampling has several advantages. First it facilitates sample rotation. Twenty percent of the sample in the NASS frame is rotated each year. Second, it allows estimation of year to year change from the 80 percent of the sample that did not change. Third, it simplifies the process of adjusting sample sizes to improve sampling efficiency.

The PSUs selected in the sample selection program are then located and further broken down into segments. The CASS system is used for this procedure as well. Just as the PSUs were originally delineated, so are the segments within each chosen PSU. Segments must be constructed using permanent boundaries, contain similar amounts of cultivation, and be equally sized. The CASS system randomly chooses one of the segments to sample from each PSU.

For data collection, segment boundaries are transferred to large scale NAPP photography. This process is completed by hand.

Data Collection and Estimation

Data Collection

NASS conducts a major area frame survey once each year in June. It is called the June Agricultural Survey (JAS). Approximately 15,000 segments are enumerated and yield approximately 50,000 farm establishments. These segments account for roughly 0.8 percent of the total land area of the 48 conterminous states. The survey produces estimates of crop acreages, grain stocks, number of and land in farms, livestock inventories, farm labor, and cash receipts at state, regional, and national levels. Importantly, the information collected during this survey provides a database of information about the farm establishments sampled through the area frame. This information is used as a sampling base for follow-on surveys for the remainder of the year. In particular, farm establishments are checked against the NASS list sampling frame to measure the incompleteness of that frame. The follow-on surveys generally use multiple-frame methodology, incorporating list samples with an area sample which account for this incompleteness.

Prior to the JAS data collection period, newly rotated segments along with residential, commercial, and non-agricultural segments are screened for the presence of farm establishments. States implementing a new area frame must screen all segments. Screening usually takes place in late April to early May. A questionnaire is filled out for segments which contain no agriculture.

The data collection period for the June Agricultural Survey begins June 1 and continues for two weeks. Enumerators conduct face-to-face interviews with operators of all farm establishments with land inside a segment, and account for all land within the segment. Enumerators are assigned anywhere from 8 to 15 segments to survey depending on distance between segments and the enumerator's experience level.

Estimation

In the NASS area frame, recall that segments are the population units and the second stage sampling units. The reporting units are the individual farm establishments within the segments. However, depending on the estimator that is used, these reporting units are defined somewhat differently. Sometimes the establishment reports only for its land *contained within the segment*. That part of a segment operated by a single establishment is referred to as a "tract." For other estimators, the establishment reports information for its entire operation. Other times only farm establishments whose operator lives inside the segment report information.

Three different estimators for summarizing area frame data are described below. Each has different advantages and disadvantages. Each may be used alone to estimate agricultural items, or in conjunction with a list frame to estimate for the undercoverage of that list. Variance formulations are not presented here in order to conform

to the length requirement for the paper. Formulations used for all three estimators ignore the second stage variance component because it is relatively small and because there is only one segment selected per PSU.

Closed Estimator

The closed estimator simply sums data associated with all land *within the segment boundaries*, and expands these "segment totals" to represent the population. A state level sample estimate using the closed estimator may be expressed mathematically as follows:

$$\hat{Y}_c = \sum_{i=1}^{L} \sum_{j=1}^{s_i} \sum_{k=1}^{r_{ij}} y_{ijk}$$

where

$$y_{ijk} = \begin{cases} e_{ijk} \sum_{l=1}^{f_{ijk}} t_{ijkl} & \text{if } f_{ijk} > 0, \\ 0 & \text{if } f_{ijk} = 0, \end{cases}$$

t_{ijkl} = the value of the survey item on the total *tract* acres operated for the l^{th} tract operation in the k^{th} segment, j^{th} sub-stratum, and i^{th} land-use stratum,

f_{ijk} = the number of tracts in the k^{th} segment, j^{th} sub-stratum, and i^{th} land-use stratum,

e_{ijk} = the expansion factor for the k^{th} segment in the j^{th} sub-stratum and i^{th} land-use stratum,

r_{ij} = the number of sample replicates or segments in the j^{th} sub-stratum, and i^{th} land-use stratum,

s_i = the number of sub-strata in the i^{th} land-use stratum,

L = the number of land-use strata in the state.

The closed estimator is simple and easy to use. Farm establishments report only for data within the segment boundaries. Reported data is easily verified and thus relatively free of reporting errors. The closed estimator can be very precise for estimating agricultural items such as planted acreages. However, other agricultural items, such as farm labor and cash receipts, can only be reported accurately for the entire farm establishment. A closed estimator is not reasonable for estimating such items. This approach usually requires a face-to-face interview to show the segment boundaries to the farm operator. Thus data collection costs are high.

Weighted Estimator

The weighted estimator uses entire farm data, and prorates (or weights) some portion of that data to each population unit (segment) in which the farm has land. A variety of weighting schemes are possible, the only restriction is that the sum of the weights for a farm across all population units will equal "one." NASS currently uses a ratio of "tract acres minus farmstead" to "entire farm acres minus farmstead" as its operational weight. Reported data for the entire farm is multiplied by this weight and summed to the segment level and then expanded for the entire population.

The state level sample estimate using the weighted estimator may be expressed mathematically as follows:

$$\hat{Y}_w = \sum_{i=1}^{L} \sum_{j=1}^{s_i} \sum_{k=1}^{r_{ij}} y_{ijk}$$

where

$$y_{ijk} = \begin{cases} e_{ijk} \sum_{l=1}^{f_{ijk}} a_{ijkl} z_{ijkl} & \text{if } f_{ijk} > 0, \\ 0 & \text{if } f_{ijk} = 0, \end{cases}$$

$$= \begin{cases} e_{ijk} \sum_{l=1}^{f_{ijk}} w_{ijkl} & \text{if } f_{ijk} > 0, \\ 0 & \text{if } f_{ijk} = 0, \end{cases}$$

w_{ijkl} = the weighted value of the survey item for the l^{th} operation with land in the k^{th} segment, j^{th} sub-stratum, and i^{th} land-use stratum,

a_{ijkl} = the weight for the l^{th} agricultural operation with land in the k^{th} segment, j^{th} sub-stratum, and i^{th} land-use stratum,

z_{ijkl} = the value of the survey item on the total acres operated for the l^{th} operation with land in the k^{th} segment, j^{th} sub-stratum, and i^{th} land-use stratum,

e_{ijk}, r_{ij}, s_i, L are previously defined.

The weighted estimator incorporates entire farm level data and thus can be used for any agricultural item. Once the "tract acres minus farmstead" value is established for

each operation, less expensive collection procedures are possible as face-to-face interviews are not required. NASS has found, however, that weighted estimates are often biased upward when the weight depends on whole farm acreage. Farm operators under-report farm acreage (which included cultivated plus non-cultivated land), which in turn causes the weight to be biased upward. The NASS operational weight suffers from this problem. By eliminating the farmstead in the weight calculation, NASS simplifies screening in agri-urban strata, where a farm operator may reside apart from his/her operation.

Open Estimator

The open estimator is a special case of the weighted estimator, which gives a weight of "one" to farm establishments whose operator resides within the segment, and a weight of zero otherwise. Data need only be collected from resident farm operators, thus reducing data collection costs and respondent burden. However, many disadvantages are associated with the use of the open estimator, and NASS has discontinued its use. First, the estimates are less precise than other weighted estimators. Second, farm operator residences are sometimes missed when screening segments in agri-urban areas. This causes open estimators to be biased downward. Intensive, and expensive, screening procedures are needed to make this estimator work satisfactorily.

Cost

The construction and maintenance of a national level area frame is a costly undertaking with respect to both labor and materials. When constructed and maintained on paper, the cost of labor far outweighed the cost of the materials. Many hours were required for the delineation of strata and PSUs on several different media. Additional hours were required for reviews. The use of the CASS system has shifted the relative cost of labor and materials. Many activities are now automated. Using this system the stratification of an average county takes approximately 44 staff hours. Using paper materials the same county would take approximately 105 staff hours.

The Arkansas frame was the last one constructed using paper materials. The process used approximately 10,000 staff hours ($86,000). Materials (including paper satellite imagery, photography, and maps covering the whole state, photo enlargements of selected segments) cost approximately $30,000. Thus the cost of building the frame was approximately $116,000, with 75 percent of the total for labor. The Oklahoma frame was larger and cost approximately $124,000 to complete. With CASS, however, only 35 percent of the total was for labor. The major recurring cost with CASS is the purchase of digital satellite imagery. CASS also had significant up front costs for equipment and software development. Over time, we expect labor costs to increase and the cost of digital satellite imagery to decrease, making the CASS system a truly cost effective medium for the construction of area frames.

Data collection costs are also of interest. NASS enumerates approximately 15,000 segments during the June Agriculture Survey each year. Data is collected during a two week time frame by approximately 1600 enumerators. Costs average $180 per segment. This includes enumerator training, travel, screening, and data collection.

Quality Control and Assessment

Quality control and assessment is an ongoing process within the frame construction process and throughout the useful life of that frame. The following sections discuss the process of discovering and correcting problems with individual segments, and procedures for assessing the deterioration of an older frame.

Problem Segments

Occasionally a segment is selected that can not be efficiently enumerated. These segments are termed "problem segments" and require immediate, careful attention. Problem segments are generally caused by one of two situations: 1) segment boundaries are not well defined, or 2) the segment is too large or contains too many farm establishments to enumerate accurately in a reasonable amount of time.

The first assessment of the quality of segment boundaries occurs when the boundaries are copied onto aerial photography. Because the line mapping data overlaid on the satellite imagery in the CASS system is usually older than the aerial photography, some boundaries chosen with CASS may not appear on the photography. In those cases, cartographers make small adjustments to the segment boundaries to accommodate the boundaries on the photography. On rare occasions, PSU and stratum boundaries are also adjusted. Care is used to avoid changing the number of sampling units. The second assessment occurs during data collection. If a boundary error is found at this point, the segment is adjusted prior to next year's survey.

Problems associated with the size of the segment and with the number of interviews required are usually

discovered during the initial screening. These are resolved by dividing the segment into a number of smaller parcels of land and randomly selecting one. The expansion factor for the new segment is appropriately modified.

Assessing the Deterioration of an Older Frame

Land utilization within each state is constantly changing. As a result, over time a state's area frame will contain an increasing number of segments that do not conform to their stratum's definition. This occurrence, in turn, damages the frame's ability to produce useful and accurate estimates. Frames exhibiting this characteristic are said to be "aging".

Bush describes a systematic approach to prioritize states for new frame construction. The approach consists of: 1) deciding upon objective criteria, or standards, by which to judge each frame, 2) ranking the states for each individual criteria, 3) assigning weights, or relative importance, to each criteria, and 4) using the weighted ranks to arrive at an overall ordering based on all criteria. Bush uses the following criteria in his assessment.

1) Percentage of segments meeting strata specifications. Assuming that almost all segments met their stratum definitions when the frame was new, this serves as a basic measure of stratification aging.

2) Relative importance of state to the national estimating program. A national level optimal sample allocation analysis is performed for commodities whose estimates rely heavily upon the area frame (as opposed to being estimated from the list frame of farm operators). The objective is to highlight states needing an increased sample size in order to reach national level precision goals.

3) Availability of current aerial photography. Though frame construction is now automated with the use of the CASS system, sampled segments are still delineated on large scale aerial photography and sent to the state offices for each survey. Ensuring the availability of current photography, therefore, decreases the possibility of adding non-sampling errors to the estimates.

This type of analysis is performed approximately every five years to insure that resources are used efficiently.

OTHER AREA FRAMES WITH A RURAL FOCUS

The remaining section of this paper reviews three other area sampling frames which are designed, in part, to collect information from farm establishments. These are 1) the area frame used by Statistics Canada for agricultural surveys; 2) the Environmental Monitoring and Assessment Program's hexagonal area frame; and 3) the area frame constructed for the National Resource Inventory Survey. The reviews are less detailed than the preceding one. They describe the purpose of each frame, and provide an overview of their design. The paper then compares and contrasts the four area frames from the prospective of collecting information from farm establishments.

Statistics Canada's Area Frame (As Used for Farm Establishments):

The Agriculture Division of Statistics Canada has been conducting a survey of farm establishments using various forms of area frame methodology since the early seventies. The major purpose in using the area frame is to account for the incomplete coverage of farm establishments on the list frame. During this time frame the quality of the list frame has greatly improved, requiring less dependence on the area frame. The agricultural area frame in Canada relies heavily on use of Enumeration Areas (EAs) and data from the quinquennial census.

The design and construction of area samples is being fundamentally revised in Canada. The previous approach used Census of Agricultural Enumeration Areas as the primary sampling units (PSUs) for the area frame. Enumeration Areas classified as "ag" in the Census (i.e., contain at least one farm headquarters) were subsampled in a two stage design similar to that used for the NASS frame. Using natural boundaries, selected PSUs were broken into 10 to 30 segments of about 6 to 10 square kilometers. A second stage sample of segments was then selected, usually one per PSU. Julien and Maranda (1990) and Ingram and Davidson (1983) discuss the earlier design.

Trepanier and Theberge present a detailed look at the redesign in a paper presented at this conference. It is a single stage design which uses the Universal Transverse Mercator projection to divide the country into 3 x 2 kilometer rectangles or cells. (In the west, 1 x 3 mile segments are used instead of the cells, and a completely different methodology is planned for Prince Edward Island.) The boundaries of these cells and of the Census Enumeration Areas are digitized and overlaid. A

computer proportionately distributes census data from an Enumeration Area into all cells that overlap that Area. Cells that straddle Enumeration Area boundaries are assigned data from both Areas. This process assigns measures of agricultural activity to the frame's sampling units. Cells that do not overlap agricultural Enumeration Areas are removed from the population. Likewise cells corresponding to urban and remote regions, forest and water are manually identified and removed. The remaining cells form the population of segments from which the single stage sample is drawn.

This population is stratified first on geographic location and then on a composite measure of agricultural activity. Sample allocation to major geographic regions is proportional to size. Allocation within geographic strata is proportional to the square root of size. The resulting sample consists of approximately 2000 segments. These are plotted on maps for data collection, where enumerators account for all land within the segments. Because of the lack of natural boundaries, the enumerator uses a grid to measure the area of each farm inside the segment rather than relying on the farmer's estimate. In the western part of the country the interview is even conducted over the telephone.

The Environmental Monitoring and Assessment Program's Hexagonal Area Frame

The United States Environmental Protection Agency established the Environmental Monitoring and Assessment Program (EMAP) in the late 1980s. While still in transition, this program is developing an integrated network for environmental monitoring with the following objectives: 1) to estimate, on a regional basis, the current status of and trends in the condition of the nation's ecological resources; 2) to monitor pollutant exposure and to understand the links between existing conditions and human-induced stresses; and 3) provide periodic statistical summaries to policy makers and the public. Inherent in these objectives is the need to statistically sample any land or water based ecological resource, including agricultural land. The information needed is clearly "area" based, and hence EMAP developed an area frame approach to their sample design.

A full description of the design of this area frame is contained in Overton, et al (1990). The process samples the land/water area of the conterminous United States via a grid composed of approximately 12,600 point locations, with 27 km. between points in each direction. The grid was constructed by centering a regular hexagon on the conterminous United States. The hexagon covered the targeted land area and parts of the adjacent continental shelf, southern Canada, and northern Mexico. Each side of the hexagon measured approximately 2,600 km. in length. Six equilateral triangles were constructed within the hexagon by connecting radial lines from the center to each vertex. Next, each side of the equilateral triangles were divided into 96 equal parts. Within each triangle, three sets of 95 parallel lines were constructed. Each set of parallel lines connected the 95 points on the one side of the triangle with their corresponding points on another side of the triangle. This process of constructing intersecting sets of parallel lines created the grid within the base hexagon. Further, these intersecting lines created regular hexagons around each grid point. Of the 28,000 points so constructed, 12,600 fell within the conterminous United States.

These form the baseline grid for the EMAP frame. However, the procedures easily lend themselves to creating additional grid points within specified hexagons whenever higher density sampling is required. From this grid baseline, various tiers of samples can be constructed.

Tier 1 Samples: Regular hexagons were formed using a grid point as the center, using the intersecting lines creating the grid point as radii, and forming sides so that the resulting regular hexagon has an area of approximately 40 sq. km. The 12,600 hexagons thus constructed form the first stage sample of primary sampling units (PSUs) of the EMAP area frame and are called the Tier 1 sample. This sample incorporates approximately 1/16th of the area of the United States. Landscape descriptions are made of each sampled PSU, and each PSU is then partitioned into resource units (those areas occupied by a single resource or land use class).

Tier 2 Samples: These samples are generally resource based. A specific resource is identified for study. PSUs containing that resource type are identified, and subsampled if appropriate. Details of the subsampling procedures were still in design stage when the design report was published. (Overton, et. al). Agricultural cropland is one major resource type of interest

Area Frame of the National Resource Inventory

The National Resources Inventory was last conducted in 1982, and is a comprehensive study of the United States' natural resources. This endeavor is the latest in a series of national inventories conducted by the Soil Conservation Service of the United States Department of

Agriculture, which have been conducted every 9-10 years since 1958. The 1982 Inventory was a joint effort between the Soil Conservation Service, the Statistical Laboratory at Iowa State University, and the U.S. Forest Service. The purpose of the Inventory was to provide statistically reliable data on land use, conservation treatment needs, erosion, and other conservation issues at various substate levels defined by either political or natural boundaries. Once again an area frame was developed to sample for this "area based" information.

A full description of the stratified, two stage design of this area frame is contained in USDA (1987). The universe of interest consisted of all nonfederal lands in the conterminous United States, Hawaii, Puerto Rico, and the U.S. Virgin Islands. The 3,300 counties in this geographic area served as the sampling base for the process.

Within each county the total surface area was stratified geographically, and land in some counties (where irrigation is important to agriculture) were also stratified according to broad resource and ownership conditions. Many small strata were constructed. In 34 states, the strata were 2-mile by 6-mile rectangular-shaped pieces of land corresponding to 12 sections. In states not covered by the public land survey system, the stratification was based on either latitude-longitude lines or the Universal Transverse Mercator projection. Always strata were constructed on a county by county basis.

Within each stratum, a two-stage area sample was drawn. The primary sampling unit was an area of land which forms a square, one-half mile on each side, containing 160 acres. In Western states some PSU's were 40- or 640-acre squares (the smaller units among irrigated land and the larger among large tracts of range land or forest). In the northeastern U.S., PSU's are 20 seconds of latitude by 30 seconds of longitude and range in size from 97 acres to 114 acres. In Louisiana and northern Maine, the PSUs are 1/2 kilometer squares (61.8 acres), while in Arkansas they are square kilometers of land. The number of PSU's selected in a given stratum depended on the variability of the county relative to land use and soil patterns, size of the county, and projected workload of data collectors. The entire sample consists of approximated 350,000 PSUs, which comprise a 3.5 percent sample of the nonfederal land area of the U.S.

Within each PSU, three point samples were selected. (Exceptions: two selected in 40-acre PSU, and one in Arkansas, Louisiana, and northern Maine). The process of selecting points assured both a random selection and a spread across the PSU. Soil Conservation Service employees collected data for each sample. Some information was collected for the entire PSU (such as area in farmsteads, enumeration of ponds, lakes, steams). Other information relating to soil type, land use, and erosion potential were collected at and for the point sights.

COMPARISON OF FRAMES

This final section summarizes and focuses the detail presented earlier in the paper by comparing and contrasting the four frames in terms of a) the purpose for which each was built and the universe over which it can provide inference, b) the sampling units used, and c) the stratification of the sampling units and what that says about estimation efficiency.

The purpose of the NASS area frame is to serve as a sampling base for producing agricultural statistics, both as a single frame and in multiple frame methodology. It provides complete coverage of all land area within the conterminous United States and Hawaii. The purpose of the Statistics Canada frame is almost identical to that of the NASS frame, except that it is used exclusively in the multiple frame context. The Canadian list frame has a higher coverage of farms than the NASS list frame, and therefore the area frame has less impact on the estimating program. It provides complete coverage of all Canadian provinces except Newfoundland. The focus of the EMAP frame and the NRI frame is environmental. Because the land and water used for agricultural production represent a significant portion of total natural resources of the United States, both frames can be used to target farm establishments. For the NRI frame, agricultural land is intended to be its main focus. It provides complete coverage of all nonfederal land in the conterminous United States plus Hawaii, Puerto Rico and the Virgin Islands. The EMAP frame is designed to focus on many different environmental resources. It provides complete coverage of all land area and water masses within the conterminous United States.

The basic sampling unit for the NASS frame is the segment, generally one square mile in size, which has natural boundaries and may be irregularly shaped. Statistics Canada uses rectangular cells, generally 3 x 2 kilometers in size, which were defined using the Universal Transverse Mercator projection rather than natural boundaries. In the west, they follow segment lines. The basic sampling units for the NRI frame are the PSUs and the three point samples selected within each sampled PSU. The PSUs are square areas, one-fourth square mile in size, that do not follow natural boundaries.

The EMAP frame uses 40 sq. km hexagons as the basic sampling unit. These where built using a grid system, and do not follow natural boundaries. In three of the four frames the lack of natural boundaries in defining the sampling unit causes more difficulty during data collection, and increases the chance of enumeration errors.

The NASS frame is built individually for each state, and population units are stratified by general land use categories and sub-stratified geographically within each state. It uses a two stage design with heavier sampling rates in intensive agricultural strata. This provides relatively efficient estimates of major agricultural production items. The area frame used by Statistics Canada is first stratified geographically and then by a measure of agricultural activity obtained from the Agricultural Census. It is a single stage design, and like the NASS frame, samples areas of intensive agriculture more heavily. The use of a single stage design and availability of Census data for stratification has the potential for making this frame the most efficient of the four for targeting farm establishments. The NRI frame is stratified geographically, but has no other stratification to target agricultural activity. This probably leads to some lack of efficiency in estimating agricultural items. The EMAP frame serves many different purposes so it is designed to spread samples geographically, but has no stratification. It is probably the least efficient for targeting farm establishments.

REFERENCES

Bush, Jeffrey. 1993. "Ranking the States for Area Frame Development." Staff Report Number SMD 93-01. Washington, D.C.: U. S. Department of Agriculture, National Agricultural Statistics Service.

Cotter, Jim and Jack Nealon. 1987. "Area Frame Design For Agricultural Surveys." Staff Report. Washington, D.C.: U. S. Department of Agriculture, National Agricultural Statistics Service.

Goebel, J. Jeffrey, Mark Reiser, and Roy D. Hickman. 1985. "Sampling and Estimation in the 1982 National Resources Inventory." *Proceedings of the American Statistical Association Meetings.*

Gordon, Daniel K. 1985. "An Investigation Of Thematic Mapper Satellite Imagery For Inventorying Fruit Trees In New York." Thesis presented at Cornell University.

Ingram, S. and G. Davidson. 1983. "Methods Used in Designing the National Farm Survey." *Proceeding of the Section on Survey Research Methods, American Statistical Association.*

Julien, C. and F. Maranda. 1990. "Sample Design of the 1988 National Farm Survey." *Survey Methodology* 16, 117-129.

Marx, Robert W. 1984. "Developing An Integrated Cartographic/Geographic Data Base For The United States Bureau Of The Census." Washington, D.C.: United States Department of Commerce, Bureau of the Census.

Mergerson, James W. 1989. "Area Frame Sampling: Sample Allocation." Internal Documentation. Washington, D.C.: United States Department of Agricultural, National Agricultural Statistics Service.

Nealon, John Patrick. 1984. "Review of the Multiple and Area Frame Estimators." Staff Report Number 80. Washington, D.C.: United States Department of Agriculture, Statistical Reporting Service.

Nealon, Jack. 1990. Revised. "Statistical Standard For Area Frame Problem Segments." Internal Documentation. Washington, D.C.: United States Department of Agriculture, National Agricultural Statistics Service.

Overton, W. Scott, Dennis White, and Don L. Stevens. 1990. *Design Report for EMAP.* EPA/600/3-91/053. Washington, D.C.: U. S. Environmental Protection Agency, Office of Research and Development.

Theberge, Alain, and John G. Kovar. 1993. "The Design of the Canadian Area Farm Survey." Florence: Presented at the 49th Session of the International Statistical Institute.

U.S. Department of Agriculture. 1987. *Basic Statistics 1982 National Resources Inventory.* Statistical Bulletin Number 756. Washington, DC.: U. S. Department of Agricultural, Soil Conservation Service.

U.S. Department of Agriculture, Iowa State University. 1987. *National Resource Inventory: A Guide For Users of 1982 NRI Data Files.* Unpublished report.

U.S. Department of Agriculture. 1992. "Agricultural Surveys Supervising and Editing Manual, Section 3." Internal Documentation. Washington, D.C.: National Agricultural Statistics Service.

METHODS OF SELECTING SAMPLES IN MULTIPLE SURVEYS TO REDUCE RESPONDENT BURDEN

Charles R. Perry, Jameson C. Burt and William C. Iwig, National Agricultural Statistics Service
Charles Perry, USDA/NASS, Room 305, 3251 Old Lee Highway, Fairfax, VA 22030

KEY WORDS: Respondent burden, multiple surveys, stratified design

Summary

The National Agricultural Statistics Service (NASS) surveys the United States population of farm operators numerous times each year. The list components of these surveys are conducted using independent designs, each stratified differently. By chance, NASS samples some farm operators in multiple surveys, producing a respondent burden concern. Two methods are proposed that reduce this type of respondent burden. The first method uses linear integer programming to minimize the expected respondent burden. The second method samples by any current sampling scheme, then, within classes of similar farm operations, it minimizes the number of times that NASS samples a farm operation for several surveys.

The second method reduces the number of times that a respondent is contacted twice or more within a survey year by about 70 percent. The first method will reduce this type of burden even further.

Introduction

The National Agricultural Statistics Service (NASS) surveys the United States population of farm operators numerous times each year. Some surveys are conducted quarterly, others are conducted monthly and still others are conducted annually. Each major survey uses a list dominant multiple frame design and an area frame component that accounts for that part of the population not on the list frame. The list frame components of these surveys constitute a set of independent surveys, each using a stratified simple random sample design with different strata definitions. With the current procedures some individual farm operators are sampled for numerous surveys while other farm operators with similar design characteristics are hardly sampled at all. Within the list frame component, two methods of sampling are proposed that reduce this type of respondent burden.

Historically, NASS has attempted to reduce respondent burden and also reduce variance. In 1979, Tortora and Crank considered sampling with probability inversely proportional to burden. Noting a simultaneous gain in variance with a reduction in burden, NASS chose not to sample with probability inversely proportional to burden. NASS has reduced burden on the area frame component of its surveys. There, a farmer sampled on one survey might be exempt from another survey, or farmers not key to that survey might be sampled less intensely. Statistical agencies in other countries have also approached respondent burden. For example, the Netherlands Central Bureau of Statistics does some co-ordinated or collocated sampling, ingeniously conditioning samples for one survey on previous surveys (de Ree, 1983).

Formal Description of Methods I and II

Method I is formally described by four basic tasks.

(a) Cross-classify the population by the stratifications used in the individual surveys. This produces the coarsest stratification of the population that is a substratification of each individual stratification.

(b) Proportionally allocate each of the individual stratified samples to the *substrata*. Use random assignment between substrata where necessary.

(c) Apply integer linear programming within each substratum to assign the samples to the labels of units belonging to the substratum so that the respondent burden is minimized.

(d) Randomize the labels to the units of substratum. The final assignment within each substratum is a simple random sample with respect to each of the proportionally allocated samples.

Method II is formally described by four basic tasks.

(a) Using an equal probability of selection technique within a stratum, select independent stratified samples for each survey. Notice that the equal probability of selection criterion permits efficient zonal sampling techniques on each survey within strata. Currently, within strata samples are selected systematically with records essentially in random order.

(b) Substratify the population by cross-classifying the individual farm units according to the stratifications used in the individual surveys.

(c) Randomly reassign within each substratum the samples associated with units having excess respondent burden to units having less respondent burden.

(d) Iterate the reassignment process until it minimizes the number of times that NASS samples a farm operator for several surveys in the substratum.

For both methods, define respondent burden by an index that represents the comparative burden on each individual sampling unit in the population. Each survey considered is assigned a burden value. When a sampling unit is selected for multiple surveys, the burden index may be additive or some other functional form dependent on the individual survey burden values. Consequently, each sampling configuration is assigned a unique respondent burden index.

For any reasonable respondent burden index, the first method minimizes the expected respondent burden. This follows easily from the following observations, where it is assumed that for each of the original surveys an equal probability of selection mechanism (*epsm*) is used within strata. First, from the independence of the original sample designs, it follows that for each individual unit the expected burden from the original stratified samples is equal to the expected respondent burden using proportional allocation followed by epsm sampling within substrata. Since the respondent burden over any population is the sum of the respondent burden on the individuals of the population, the equality holds for the entire population or any subpopulation including the substrata. That is, the expected respondent burden over any arbitrary substratum for the proportionally allocated samples is equal to the expected respondent burden of the original stratified sample allocations over the substratum. Originally, these allocations are random to each substratum, constrained only so that the substratum sample sizes sum to their stratum sample size. Second, for the first method the respondent burden is minimized over each substrata by the linear programming step.

Regarding variance reduction, this means that if the original sample was selected using simple random sampling within each stratum, then the first method reduces respondent burden without any offsetting increase in variance, since proportional allocation is at least as efficient as simple random sampling. However, the first method would be less efficient for variance than zonal sampling unrestricted by burden. But the second method, by reallocating some zonal sampling units to reduce respondent burden, may only slightly increase variance over no reallocation and then only when zonal sampling is effective.

A Simple Simulation of Method I

Method I reduces respondent burden in the following simulation of two surveys. Survey I samples $n = 20$ from $N = 110$. Survey II also samples $n = 20$ from $N = 110$, though each of its strata has either a larger or smaller population size ($N_{\cdot 1} = 40$ and $N_{\cdot 2} = 70$) than the corresponding strata of survey I ($N_{1\cdot} = 30$ and $N_{2\cdot} = 80$). Here, the first subscript corresponds to the first survey, with its strata 1 and 2. Similarly, the second subscript corresponds to the second survey. For example, $N_{21} = 30$ corresponds to the size of the population in stratum 2 of survey I and in stratum 1 of survey II, while $\bar{n}_{21}^{(1)} = 3.75$ corresponds to the proportional allocation of survey I's stratum 2 sample, $n_{2\cdot}^{(1)} = 10$, to the population in both stratum 2 of survey I and stratum 1 of survey II.

Survey I	Survey II: Stratum 1	Survey II: Stratum 2	
Stratum 1	$N_{11} = 10$ $\bar{n}_{11}^{(1)} = 3.33$ $\bar{n}_{11}^{(2)} = 2.5$	$N_{12} = 20$ $\bar{n}_{12}^{(1)} = 6.67$ $\bar{n}_{12}^{(2)} = 2.85$	$N_{1\cdot} = 30$ $n_{1\cdot}^{(1)} = 10$
Stratum 2	$N_{21} = 30$ $\bar{n}_{21}^{(1)} = 3.75$ $\bar{n}_{21}^{(2)} = 7.5$	$N_{22} = 50$ $\bar{n}_{22}^{(1)} = 6.25$ $\bar{n}_{22}^{(2)} = 7.15$	$N_{2\cdot} = 80$ $n_{2\cdot}^{(1)} = 10$
	$N_{\cdot 1} = 40$ $n_{\cdot 1}^{(2)} = 10$	$N_{\cdot 2} = 70$ $n_{\cdot 2}^{(2)} = 10$	

With two surveys, at most we will sample a respondent twice. For the above two surveys, without any proportional allocation, we simulated two independent stratified simple random samples 3 million times. These simulated samples produced, on average, 3.6 double hits for the whole population of 110 potential respondents, and four percent of the simulations produced 7 or more double hits. With the proportional allocations indicated in the diagram for Method I, the population exceeds the total sample for both surveys in each substratum, so no sampling unit needs to be selected for both surveys. The high respondent burdens of independent sampling are reduced to 0 double hits with Method I!

Operational Description

Basic Notation

Let $\mathcal{U} = \{u_i\}_{i=1}^{N}$ be a finite population of size N. Suppose that $\mathcal{U}$ is surveyed on K occasions and that on each occasion a different independent stratified design is used. For these K stratified designs, denote the survey occasion by $k = 1, 2, \ldots K$ and let us use the following notation.

$H^{(k)}$:the number of strata for design k,

$\mathcal{U}_h^{(k)}$:the units (the set of them) in stratum h for design k,

$N_h^{(k)}$:the size of stratum h for design k,

$n_h^{(k)}$:the sample size in stratum h for design k,

$f_h^{(k)} = n_h^{(k)}/N_h^{(k)}$:the sampling fraction in stratum h for design k,

$n^{(k)} = \sum_{h=1}^{H^{(k)}} n_h^{(k)}$:the overall sample size for design k, and

$N = N^{(k)} = \sum_{h=1}^{H^{(k)}} N_h^{(k)}$:the overall population size.

Remark

Requiring the population to be exactly the same for each survey may seem rather restrictive. However it is not, since, for each survey, one can easily introduce an extra stratum that contains the units not covered by that survey. Obviously the sample sizes associated with the extra noncovered strata are taken to be zero. This permits one to apply either Method I or Method II over years.

Warning: In multiyear applications, care must be taken to ensure that no information from the sample data is used to update any of the frames being considered. Failure to do so can lead to biased estimates. These are the same restrictions that apply to the permanent random number techniques discussed by Ohlsson (1993).

Method I

Using this notation for Method I, we next describe a sequence of simple data manipulation steps that can be used operationally to perform tasks (a) through (d) on page 1 for each of the K surveys.

Suppose that each unit, u_i, of the population $\mathcal{U}$ has been stratified for each of the K surveys. Further suppose that this information has been entered into a file containing N records, so that the ith record contains the stratification information for unit i. To be definitive, assume that the variable $S(k)$ denotes the stratum classification code for survey k and that $S(k : i)$ denotes the value of the stratum classification code for unit u_i.

For each survey k ($k = 1, 2, \ldots, K$) perform the following sequence of operations.

(a) Sort the data file by the variables $S(k), \ldots, S(K), S(1), \ldots, S(k-1)$. This will hierarchically arrange the records of the population, first by the stratification of survey k, by the stratification of survey $k+1$ within the stratification of survey k, then by the stratification of survey $k+2$ within the stratification of survey $k+1$, ..., by the stratification of survey K within the stratification of survey $K-1$, then by the stratification of survey 1 within the stratification of survey K, ..., then by the stratification of survey $k-1$ within the stratification of survey $k-2$. In terms of the *substrata* formed by the cross-classification, the records of the population are arranged sequentially after sorting as

$$\mathcal{U}^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{1,\ 1\ ,\ldots,1,1,\ldots,\ 1\ ,\ 1},$$

$$\mathcal{U}^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{1,\ 1\ ,\ldots,1,1,\ldots,\ 1\ ,\ 2},$$

$$\vdots$$

$$\mathcal{U}^{(k,k+1,\ldots,K,1,\ldots,k-2,\ k-1\ )}_{1,\ 1\ ,\ldots,1,1,\ldots,\ 1\ ,H^{(k-1)}},$$

$$\mathcal{U}^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{1,\ 1\ ,\ldots,1,1,\ldots,\ 2\ ,\ 1},$$

$$\vdots$$

$$\mathcal{U}^{(k\ ,\ k+1\ ,\ldots,\ K\ ,\ 1\ ,\ldots,\ k-2\ ,\ k-1\ )}_{H^{(k)},H^{(k+1)},\ldots,H^{(K)},H^{(1)},\ldots,H^{(k-2)},H^{(k-1)}-1},$$

$$\mathcal{U}^{(k\ ,\ k+1\ ,\ldots,\ K\ ,\ 1\ ,\ldots,\ k-2\ ,\ k-1\ )}_{H^{(k)},H^{(k+1)},\ldots,H^{(K)},H^{(1)},\ldots,H^{(k-2)},H^{(k-1)}}$$

where

$$
\begin{aligned}
& \mathcal{U}^{(k,\ldots,K,1,\ldots,k-1)}_{h_k,\ldots,h_K,h_1,\ldots,h_{k-1}} \\
&= \mathcal{U}^{(k)}_{h_k} \bigcap \ldots \bigcap \mathcal{U}^{(K)}_{h_K} \bigcap \mathcal{U}^{(1)}_{h_1} \bigcap \ldots \bigcap \mathcal{U}^{(k-1)}_{h_{k-1}} \\
&= \mathcal{U}^{(1)}_{h_1} \bigcap \ldots \bigcap \mathcal{U}^{(k-1)}_{h_{k-1}} \bigcap \mathcal{U}^{(k)}_{h_k} \bigcap \ldots \bigcap \mathcal{U}^{(K)}_{h_K} \\
&= \mathcal{U}^{(1,2,\ldots,k-1,k,k+1,\ldots,K)}_{h_1,h_2,\ldots,h_{k-1},h_k,h_{k+1},\ldots,h_K}
\end{aligned}
$$

Both the size and sequential arrangement of the substrata of stratum h for survey k are displayed schematically as

$$
\boxed{N^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{h,\ 1\ ,\ldots,1,1,\ldots,\ 1\ ,\ 1}}
$$

$$
\boxed{N^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{h,\ 1\ ,\ldots,1,1,\ldots,\ 1\ ,\ 2}}
$$

$$
\vdots
$$

$$
\boxed{N^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{h,h_{k+1},\ldots,h_K,h_1,\ldots,h_{k-2},h_{k-1}}}
$$

$$
\vdots
$$

$$
\boxed{N^{(k,\ k+1\ ,\ldots,\ K\ ,\ 1\ ,\ldots,\ k\text{-}2\ ,\ k\text{-}1\ )}_{h,H^{(k+1)},\ldots,H^{(K)},H^{(1)},\ldots,H^{(k-1)}_{-2},H^{(k-1)}_{-1}}}
$$

$$
\boxed{N^{(k,\ k+1\ ,\ldots,\ K\ ,\ 1\ ,\ldots,\ k\text{-}2\ ,\ k\text{-}1\ )}_{h,H^{(k+1)},\ldots,H^{(K)},H^{(1)},\ldots,H^{(k-1)}_{-1},H^{(k-1)}}}
$$

where

$$
N^{(k,\ldots,K,1,\ldots,k-1)}_{h_k,\ldots,h_K,1,\ldots,H_{k-1}}
$$

denotes the number of units in

$$
\mathcal{U}^{(k,\ldots,K,1,\ldots,k-1)}_{h_k,\ldots,h_K,1,\ldots,H_{k-1}} .
$$

(b) To randomly proportion the sample $n^{(k)}_h$ for stratum h of survey k to the subintervals of stratum k:

(1) Divide the length of stratum h for survey k, $N^{(k)}_h$, into a sequence of $n^{(k)}_h$ subintervals of integer length that differ in length by at most 1. Do this by forming $\frac{N^{(k)}_h}{n^{(k)}_h}$ as yet unpopulated subintervals, each with the length $n^{(k)}_h$, leaving $N^{(k)}_h - \left(\left[\frac{N^{(k)}_h}{n^{(k)}_h}\right] n^{(k)}_h\right)$ imaginary population units to be assigned. Randomly distribute these remaining imaginary units (without replacement) to the $\left[\frac{N^{(k)}_h}{n^{(k)}_h}\right]$ subintervals. Now populate these subintervals by randomly selecting a starting unit from the $N^{(k)}_h$ units. This starting unit begins the first subinterval, with its size randomly determined as above, $\left[\frac{N^{(k)}_h}{n^{(k)}_h}\right]$ or $\left[\frac{N^{(k)}_h}{n^{(k)}_h}\right] + 1$. Sequentially continue to populate the above subintervals, wrapping around to the first unit for one of the subintervals. This method of forming subintervals will let us keep the same probability of selection $\frac{n^{(k)}_h}{N^{(k)}_h}$ for each unit in that subinterval. It does not choose a sample.

(2) Randomly select an integer from each subinterval [while this integer corresponds to a population unit, it is not used here to select that population unit–for that, see (d) below].

The number of these random integers falling in the interval corresponding to

$$
N^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{h_k,h_{k+1},\ldots,h_K,h_1,\ldots,h_{k-2},h_{k-1}}
$$

in the sequential ordering is the size of the randomly proportioned sample for survey k to be drawn from the substratum population

$$
\mathcal{U}^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{h_k,h_{k+1},\ldots,h_K,h_1,\ldots,h_{k-2},h_{k-1}} .
$$

Denote this sample size for the substratum by

$$
m^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{h_k,h_{k+1},\ldots,h_K,h_1,\ldots,h_{k-2},h_{k-1}}
$$

or

$$
m^{(k)}_{h_1,\ldots,h_k,h_{k+1},\ldots,h_K}
$$

where the subscripts in the last expression are understood to be in natural order.

Repeating steps (a) and (b) above for each of the K surveys, we have randomly proportioned the K original stratified sample sizes to the substrata.

(c) Next we describe how to use integer linear programming to assign *within a substratum* the above proportioned samples to the substratum unit labels–not specific population units yet. We do this so that the respondent burden is minimized for an *arbitrary* positive linear respondent burden function (index).

Suppose that $m^{(1)}, m^{(2)}, \ldots, m^{(K)}$ samples have been randomly proportioned to a substratum of size M. Clearly the random proportioning procedure described above insures that $m^{(k)} \leq M$ for $k = 1, 2, \ldots, K$.

Moreover, if the size of the total sample $m = m^{(1)}+m^{(2)}+\ldots+m^{(K)}$ randomly proportioned to the substratum is less than or equal to M, then any positive linear respondent burden index is minimized by selecting the total sample m by simple random sampling (SRS) without replacement (WOR) where the first m_1 units selected are associated with survey I, the second m_2 units selected are associated with survey II, etc.

If the size of the total sample $m = m^{(1)} + m^{(2)} + \cdots + m^{(K)}$ is greater than M, then linear integer programming can be used to find an assignment of the total sample to the (unspecified) *labels* of the stratum that minimizes the respondent burden. Reiterating, we are working with labels here, so we are considering the burden of an arbitrary unit in the substratum, not the population units themselves, though we will use the natural terminology "population unit." When assigning samples from K surveys to the population units, there are 2^K possible ways of assigning the samples to any one population unit. These possible assignments can be represented by the 2^K K-dimensional vectors, call them *assignment configurations*,

$$\vec{v}_1 = \begin{pmatrix}0\\0\\0\\\vdots\\0\end{pmatrix}, \quad \vec{v}_2 = \begin{pmatrix}1\\0\\0\\\vdots\\0\end{pmatrix}, \quad \vec{v}_3 = \begin{pmatrix}0\\1\\0\\\vdots\\0\end{pmatrix},$$

$$\vdots$$

$$\vec{v}_{K+1} = \begin{pmatrix}0\\0\\0\\0\\\vdots\\1\end{pmatrix}, \vec{v}_{K+2} = \begin{pmatrix}1\\1\\0\\0\\\vdots\\0\end{pmatrix}, \vec{v}_{K+3} = \begin{pmatrix}1\\0\\1\\0\\\vdots\\0\end{pmatrix},$$

$$\vdots$$

$$\vec{v}_{2^K-1} = \begin{pmatrix}0\\1\\1\\1\\\vdots\\1\end{pmatrix}, \quad \vec{v}_{2^K} = \begin{pmatrix}1\\1\\1\\1\\\vdots\\1\end{pmatrix}$$

where component k of the vector is 1 if the unit is sampled for the kth survey and 0 otherwise. Now we must determine the number x_1 of the population units to assign the configuration $\vec{v}_1$, the number x_2 to assign the configuration $\vec{v}_2$, ..., the number x_{2^K} to assign the configuration $\vec{v}_{2^K}$.

Suppose the ith assignment configuration, represented by the ith assignment configuration vector $\vec{v}_i$, produces a respondent burden of $b_i \geq 0$. Then the problem of assigning the $m^{(1)}, m^{(2)}, \ldots, m^{(K)}$ samples to the M (unspecified) unit labels such that the total respondent burden over the substratum is minimized is equivalent to minimizing the linear objective function (respondent burden index)

$$\begin{aligned} & f(x_1, x_2, \ldots, x_{2^K}) \\ = \; & b_1x_1 + b_2x_2 + \ldots + b_{2^K}x_{2^K} \\ = \; & \vec{b}\,\vec{x}^T \end{aligned}$$

subject to the $K+1$ linear constraints

$$\begin{cases} \vec{v}_1x_1 + \vec{v}_2x_2 + \cdots + \vec{v}_{2^K}x_{2^K} = \vec{m} = \\ \quad (m^{(1)}, m^{(2)}, \ldots, m^{(K)})' \quad \leftarrow \text{K constraints} \\ x_1 + x_2 + \cdots x_{2^K} = M\,, \end{cases}$$

where $x_1, x_2, \ldots, x_{2^K}$ are non-negative integers.

Since $\vec{v}_{K+2}, \ldots, \vec{v}_{2^K}$ can each be written as a nonnegative integer combination of $\vec{v}_2, \ldots, \vec{v}_{K+1}$ and since $m^{(k)} \leq M$ for each k, it is easy to see that

$$\begin{aligned} & \vec{v}_2x_2 + \vec{v}_3x_3 + \cdots + \vec{v}_{2^K}x_{2^K} \\ = \; & (m^{(2)}, m^{(3)}, \ldots, m^{(K)})' \end{aligned}$$

has a solution over the nonnegative integers, say $x_2, \ldots, x_{2^K}$. Setting

$$x_1 = M - x_2 - x_3 - \cdots - x_{2^K}$$

then provides a feasible solution to the integer linear programming problem. So there exists a solution and hence there exists an optimal solution.

(d) Finally, select specific sampling units u_i from the population. Consider a specific substratum and treat other substrata similarly. From the results of (c) above, we now randomly choose x_2 farmers from the M substratum farmers for the configuration $\vec{v}_2$, randomly choose x_3 farmers for the configuration $\vec{v}_3, \ldots$, randomly choose x_{2^K} farmers for the configuration $\vec{v}_{2^K}$. This sample of farmers reduces burden, yet within each stratum of each survey, this approach selects farmers with equal probability.

Note that this sample is not a type of systematic sample–the randomness in (b)-(2) reveals this.

Method II

In Method II, a sample is selected by some preferred technique. That sample might be selected by some equal probability of selection technique using zonal sampling to reduce variance, eg, by Chromy's Procedure, Chromy (1981). Method II largely retains that sample, but alters it to reduce burden. Thus Method II alters the sample by redistributing it within the substrata.

Since this Method II is no more complicated than Method I and has many similarities to it, the following description is brief.

(a) Within each stratum of each of the K surveys, independently select a sample with equal probability.

(b) Cross-classify the population as in (a) of Method I. This not only cross-classifies the population, it also cross-classifies the sample chosen in (a) of Method II. From the

$$N^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{h_k,h_{k+1},\ldots,h_K,h_1,\ldots,h_{k-2},h_{k-1}}$$

units in the substratum population

$$U^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{h_k,h_{k+1},\ldots,h_K,h_1,\ldots,h_{k-2},h_{k-1}},$$

denote the number sampled by

$$m^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{h_k,h_{k+1},\ldots,h_K,h_1,\ldots,h_{k-2},h_{k-1}}.$$

This subsample size will not be changed, but it will be distributed among the substratum's population in (c) below.

(c) Within a substratum, reassign or swap some of the surveys associated with a sampling unit having excess respondent burden to a sampling unit have less respondent burden. If the respondent burden index is linear, then only one survey for one sampling unit need be reassigned to reduce burden. For example, when we measure respondent burden by the number of times we hit a farmer with a survey. Then we would move one survey from the farmer who got 4 hits to the farmer who got 0 hits, or to the farmer who got 2 hits if no farmer got 0 or 1 hit.

If the respondent burden is non-linear, then sometimes more than one survey must be reassigned to reduce burden. And when respondent burden in non-linear, then sometimes three sampling units (not two) must swap to ever reduce burden.

(d) Repeat (c) above until no reassignments can be made. Then respondent burden has been minimized.

With this method, one might want to retain most of the original sample selection for the first survey but not necessarily for the other surveys. Then, in (c), try to reassign other surveys before reassigning the first survey. Sequential application of Method II is justified since each survey uses equal probability of selection in each stratum which implies that all units of a substratum have the same selection probability for any given assignment configuration.

Some NASS Examples

NASS administers many surveys with a large number of strata. For example, the Farm Costs and Returns Survey (FCRS/COPS) may have 18 strata, the Agriculture Survey may have 17 strata, and the Labor Survey may have 8 strata. This many strata over many surveys brings skepticism to any use of Methods I or II. One would expect many combinations of strata to contain but one individual, even for three surveys. Methods I and II could never reduce burden on such a sparsely (one individual) populated combination of strata. Fortunately, most stratum combinations are empty while other combinations are well populated.

Indeed, not only are many substratum combinations empty, many survey sampling combinations are empty. In some initial testing over nine major surveys, only 58 of the $2^9 = 512$ possible survey combinations occurred in Kansas and only 62 in Arkansas based on 1991 data. This fortuitous limitation on survey combinations gives some optimism that many combinations of strata will be well populated. A look at the number of population units selected for multiple surveys provides further optimism (see Table 1).

No burden exceeds five surveys. No sampling unit was selected for more than five surveys, indicating that the possible number of substrata with only one unit is limited somewhat.

There is some optimal combination of surveys to consider when reducing respondent burden by either Methods I or II. More surveys result in too few farmers being classified for any of the many substrata combinations. Fewer surveys prevent

Table 1: Number of Survey Hits over Nine Surveys in 1991

Hits	Arkansas Frequency	Kansas Frequency
0	3491	21474
1	21125	40900
2	6136	8638
3	846	938
4	60	74
5	1	7

Methods I and II from reducing any large burdens on some farmers; eg, when NASS surveys one farmer for five different surveys.

In 1991, for the four major surveys – FCRS, Labor, Quarterly AG, and Cattle/Sheep – NASS initially sampled the following numbers of farmers.

Survey	Arkansas	Iowa	Kansas
FCRS	666	1836	1356
Labor	576	728	440
Quarterly AG	4442	6477	5881
Cattle/Sheep	1727	5507	3204

Method II reduced burden by about 70 percent over the three states Arkansas, Iowa and Kansas in 1991 and 1992. Table 2 below summarizes these reductions of burden. Since the NASS samples were essentially random within strata, a huge reduction can be made in burden with no cost (increase) in variance.

Table 2: Reduction in Multiple Sample Selections Using Method II for the FCRS, Labor, Quarterly AG, and Cattle/Sheep Surveys

Number Selections	1991		
	Current	Method II	% Reduction
4	0	0	–
3	159	50	69
2	2620	782	70
Total	2779	832	70
Arkansas	733	205	72
Iowa	1105	252	77
Kansas	941	375	60

Number Selections	1992		
	Current	Method II	% Reduction
4	6	4	33
3	112	28	75
2	2371	749	68
Total	2489	781	69
Arkansas	735	124	83
Iowa	801	204	75
Kansas	953	453	52

Acknowledgements

The authors would like to express their appreciation to George Hanuschak for assistance in initiating and supporting this project early on. They would also like to thank Ron Bosecker and Jim Davies for continued support.

State Statistical Offices are quite concerned about individual farmer respondent burden and want the Agency to pursue methods to reduce it while maintaining an acceptable level of statistical integrity. This report is the first such statistical attempt in recent years in the Agency. Special thanks go to State Statisticians T. J. Byran, Ben Klugh, Dave Frank and Howard Holden for providing their substantial insights into the definition and potential relief of individual farmer respondent burden.

References

Chromy, James R. (1981). "Variance Estimators for a Sequential Sample Selection Procedure," *Current Topics in Survey Sampling*, D. Krewski, R. Platek, and J.N.K. Rao (Eds). Academic.

de Ree, S.J.M. (1983). *A System of Co-ordinated Sampling to Spread Response Burden of Enterprises.* Netherlands Central Bureau of Statistics.

Ohlsson, Esbjörn (1993). "Coordination of Samples Using Permanent Random Numbers," ASA International Conference on Establishment Survey Proceedings.

Tortora, Robert D. and Crank, Keith N. (1978). *The Use of Unequal Probability Sampling to Reduce Respondent Burden.* USDA/National Agricultural Research Service.

ENSURING QUALITY IN U.S. AGRICULTURAL LIST FRAMES[1]

Cynthia Z.F. Clark, National Agricultural Statistics Service, Elizabeth Ann Vacca, Census Bureau
Cynthia Z.F. Clark, NASS, 14th & Independence Ave., Washington, D.C.

KEY WORDS: Business registers, record linkage, census and survey list frames

ABSTRACT

In the United States, agricultural data are collected by the Bureau of the Census in the Department of Commerce, and the National Agricultural Statistics Service in the Department of Agriculture. Both agencies are mandated by law to collect data on the agricultural economy. Title 13 of the United States Code (USC) requires the Census Bureau to conduct a quinquennial census of agriculture, providing detailed data on agricultural operations for each county. USC Title 7 requires the National Agricultural Statistics Service to collect data on agriculture and its market. The census of agriculture has mandatory data reporting requirements whereas surveys conducted by the NASS have voluntary participation.

Both agencies are required to protect the confidentiality of the individual record data but are subject to different legal requirements. Title 13 restricts access to census data to sworn officials and census employees, prohibiting the release of the census of agriculture list or census data to the NASS, or to any other organization. Title 7 permits the use of data collected by NASS for statistical purposes, and thus enables the Census Bureau to use the NASS list in building the census list. The NASS list, however, does not have complete coverage of the universe of farm operations. Thus, it is necessary for each agency to compile its own list frame to meet its mandated statistical needs.

Both U.S. agencies use the same definition of a farm - a place from which $1,000 or more of agricultural products were sold during a calendar year. A new census list is created every five years by linking agricultural statistical, administrative, and commodity lists of establishments. The NASS list was developed in the late 1970's from similar types of lists and is continually maintained. The overall quality of both census and survey data is dependent upon the quality of the list frames. The paper focuses on five attributes of list frames - scope, accuracy, duplication, coverage, and cost efficiency - relating to quality. List development procedures and quality indicators are discussed and compared.

1. AGRICULTURE CENSUS LIST FRAME

Since 1969 the census of agriculture has been conducted using a mail-out/mail-back data collection procedure in lieu of personal enumeration. Prior to each census, the Census Bureau assembles records of individuals, businesses, and organizations identified as having some association with agriculture. This includes files from the previous census, administrative records of the Internal Revenue Service (IRS) and the Social Security Administration (SSA), and statistical records of NASS. Additionally, lists are obtained for specialized operations (e.g. nurseries and greenhouses, specialty crop farms, poultry farms, fish farms, livestock farms, cattle feedlot operations, grazing permittees) from State and Federal government agencies, trade associations, and similar organizations. Lists of companies having multiple establishments (or locations) producing agricultural products are obtained from previous censuses and updated using the information from the Standard Statistical Establishment List maintained by the Census Bureau.

After the various address lists are acquired, the Census Bureau performs record linkage to remove duplicate addresses, screens for nonfarm records, and prepares mail labels for each address. Five major operations are required for record linkage: format and standardization, business or personal identification number linkage (Employer Identification Number, EIN, or Social Security Number, SSN), geographic coding and ZIP code verification, alphabetic name linkage, and clerical review of potential duplicate record sets. For the past four censuses, two similar phases of address linkage were conducted to permit incorporating more current addresses than would have been available using a single linkage.

The format and standardization operation places each source record into a common format; edits the source records; and assigns name control,

[1]This paper reports the general results of research undertaken by the staff of the Census Bureau and the National Agricultural Statistics Service. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau or the NASS.

processing codes and size codes to each record (Gaulden, 1990). The name control uses the first four characters of the primary name, containing a minimal of three non-numeric characters, not appearing in a special agricultural "skip list" dictionary. This dictionary contains over 1,000 words and abbreviations (such as Farm, Dairy, Bros) which could conceivably appear in the name field but were not likely to be the surname.

Processing codes facilitate the use of the most reliable information in the final list record. Initially, each record is assigned a name and address priority code based on the expected currency of the address information of the record source. During record linkage, the name and address of the record within a linked record set with the highest priority is retained. Each record is also assigned a size code based on the estimated total value of agricultural products (TVP) to be sold in the census year derived from agricultural data on the source record. During record linkage, the size code from each record in linked record sets is transferred from each of the deleted duplicate records to the retained record. This allowed for the derivation of both a "source combination code" indicating all the sources for the final record and a "final size code" based on the reliability of size information for each source.

Records are then linked by computer on EIN and SSN. All remaining unmatched and potential duplicate records have the addresses coded by geographic area, the names identified according to part, and the names coded using SOUNDEX procedures (where vowels and double consonants are deleted from names). The geographic coding system was designed to ensure that all records contained standardized and edited geographic codes, prior to record linkage. The name and address linkage procedure adapted the Fellegi-Sunter probability model relying on the extent of agreement between name, address, and other record identifiers to determine duplication within ZIP code (or area) blocks. An additional linkage of historically identified duplicate records is conducted. Duplicate records are deleted, retaining the record whose source is deemed to have the highest quality address information. Potential duplicate records are reviewed clerically to determine which duplicate records to delete.

2. THE NASS LIST FRAME

The NASS conducts ongoing monthly, quarterly, and annual probability based sample surveys of the agricultural sector. The samples are selected from the list frame and a land-based area frame. The list frame consists of a frame for each state that is maintained and updated by each NASS State Office using procedures developed by a NASS headquarters unit. The NASS list frame is designed to provide good coverage of large and commodity specific operations. Because NASS surveys use dual frame estimates derived from the list frame and a complete area frame, it is not necessary to achieve farm universe coverage with the list frame.

Prior to the implementation of probability surveys, NASS maintained "reporter lists" of farm operators who were willing to provide their agricultural data to the state agricultural statistical office. The quality of these state lists varied considerably. In the mid-70's the agency undertook a large effort to develop a list that would provide the frame for sample surveys with standard address and agricultural data across states for each farm operation record. To build the new NASS list, each state secured lists of potential agricultural operations to use in conjunction with the state reporter list. Lists were selected based on a composite source evaluation that weighted such factors as the degree of coverage, the frequency of updates, the type of update procedure, the list medium, the identifiers on the list, the agricultural data on the list, the use of the list, the number of records, and the cost. Lists came from such sources as the Agricultural Stabilization and Conservation Service (ASCS), the Rural Electrification Administration, and agricultural trade associations. However, regulations prohibited the release of tax records of the IRS and farm employer records of the SSA to NASS.

Similar to the census procedures, the records for each state were formatted and linked removing duplicate records. The linkage system first matched on SSN, EIN, telephone number, and two other identifiers. Surnames were coded using the New York State Identification and Intelligence Service (NYSIIS) procedure, a coding procedure that NASS research found superior to SOUNDEX for blocking (Lynch). Place names were used to assign latitude and longitude to each record for construction of a distance function for matching place names within the NYSIIS blocks. The linkage system used an adaption of the Fellegi-Sunter model that relied on the frequency of occurrence of name, address, and identifier components, and specified types of errors within the file, to determine duplication within NYSIIS blocks for each ownership type. Records determined to be probable duplicates were linked with one record being placed on the frame. An extensive clerical review of the identified potential duplicate records was conducted and the duplicates were removed from the list. Agricultural data (referred to as control data), compiled from existing farm records and

special surveys, was appended to the address record to use for sample stratification.

Each NASS state office has the responsibility for maintaining the frame for its state. Each office determines how it will update name and address information and control data, add records for new farm operators, and identify records of individuals that no longer have farm operations. The headquarters office supports the state offices by developing and maintaining computer programs and procedures to assist in this process. An interactive system has been developed for updating record information in each office. Each state prepares an annual plan for list frame development and maintenance taking into account evaluation of the state list frame and state specific survey needs. The headquarters list frame group reviews, advises, and supports the plan.

When a state wants to add ASCS or other source records to its state list frame, a match is first conducted using SSN and EIN (if available on the records). Matched records are retained with the NASS name and address and control data from all sources. At the option of each state office, nonmatching records can be run through a resolution program. This program replicates the Fellegi-Sunter process used in the original linkage with the exception that records identified as probable matches are retained as separate records on the frame unless clerically deleted. The resulting nonmatching records are added to the state list as inactive records. These inactive records become the source for "criteria surveys" (described in Section 3.3), clerical review, or personal enumeration to determine their farm status. Records for valid farm operators are placed on the active farm frame.

3. SCOPE OF THE LIST FRAMES

Both agencies face the challenge of determining the scope of the list frame by focusing on its size and composition. The size of the frame impacts the cost of data collection and processing, and the respondent burden (directly for the census and indirectly, through the efficiency of the sample design, for NASS). The frame composition affects the resulting quality of the survey data. Records from the same source often have common characteristics impacting record quality. If the frame contains records that are thought to represent farm operations but, in fact, do not have an associated operation, the integrity of the resulting data is compromised. Nonfarm operators often do not respond to a request for agricultural information. The statistical procedures for either a census or survey need to adjust for such nonresponse.

Response to the 1987 Census of Agriculture illustrates the impact of the scope of the frame on the census operation. Of the 83.1 percent responding to the 1987 census, 54 percent were farm operators, 44 percent were not farm operator addresses, and 2 percent were unclassifiable. Of the 13.3 percent not responding to the census an estimated 44 percent were farm operators. A sample survey with its corresponding sampling variability was used to account for the nonresponding farm operations in the final census data. The remaining 3.6 percent had undeliverable addresses, adding to the operation costs. This section discusses how the two organizations determine the scope of their respective frames.

3.1 Size of the List Frames

For the census, the final list becomes the address list for the mail enumeration. All records not retained in the list are placed in inactive files and are excluded from data collection. For the NASS, the final list is the sampling frame for surveys, using agricultural data residing on each record for stratification.

Cost and burden considerations severely limited the size of the 1992 census mailings -- 3.55 million records. The best strategy for list compilation with such constraints appeared to be to limit the number of less reliable input addresses, resulting in 12.4 million source records. After linking records identified as duplicates by name, address, SSN, and EIN, 4.9 million records remained at the end of the second linkage phase. Of these records, approximately 1.1 million came exclusively from nonfarm past census or NASS sources and were deleted from the mail list. Statistical classification analysis was effectively used to identify 230,000 records as least likely to represent farm operations and candidates for potential removal from the list. After minor modifications, the final census list contained 3.55 million records.

The 1992 NASS list frame consisted of 1.74 million records that were active farms or agribusinesses, a three percent increase over 1991. Additionally, there were 1.66 million inactive records. In 1991, there were 763,930 inactive records on the frame, known to be either nonfarms or out-of-business agricultural establishments. Some of these records came from new sources and their farm status had not yet been resolved; others were being retained on the list as nonfarms to aid in future identification of the record farm status during the linkage and resolution process.

3.2 Sources of List Addresses

Because of the census focus, concerted efforts

are made to include in the list compilation all important sources of agricultural record information. A two stage linkage process is used to permit the incorporation of IRS records from the two years prior to the census. However no IRS births are picked up for the year of the census. Two notable changes in list sources and their content were made for the previous two censuses. For these censuses, the Census Bureau used the agricultural farm operation list of the NASS for all 50 states as contrasted with the 31 states available for use in the 1982 census list compilation. The NASS list, provided to the Census Bureau prior to each linkage stage, enables the census list to include new records and updates to existing records as of April of the census year.

In 1992, the Census Bureau did not use all ASCS records directly as had been done in the 1978 and 1982 censuses. This decision was based on the expected inclusion of valid ASCS farms into the NASS lists, the ASCS list size, and the reliability of the ASCS name and addresses. When the ASCS records were used as a census source many ASCS records did not match other source records and required screening (25 percent of the 3.0 million records in the 1982 Farm and Ranch Identification Survey were obtained from ASCS only). Many of the addressees were determined to be landlords rather than farm operators. Of the 1982 census farm respondents, only 1.4 percent came uniquely from the ASCS list.

Following the 1987 census, tabulations of the mail list by address sources provided the contribution of each source to the list. The percent of enumerated census farm records appearing on only one source was 14.9 for IRS, 3.9 for the previous census, and 3.3 for NASS. The percent of unique farms among the respondents was much higher from the IRS list (10.8) and the special lists (8.7) than for the other source lists (3.9 for 82 census list and 4.1 for NASS list).

The NASS primarily uses information from its own surveys and data collections to update record identifiers and data. It obtains specialized lists from trade associations and other agricultural organizations to extend list frame coverage. These are as diverse as lists of farm vehicles, general livestock, cattle, hogs, sheep, equine, field crops, poultry producers, pesticide applicators, fruit growers, vegetable growers, nut tree producers, bee and honey producers, floriculture and nurseries, wool producers, agricultural labor employers, and other specialty commodity producers. Since the agency received funding for the Pesticide Data Program and the Water Quality Program in the early 90's and began surveys of chemical usage on vegetable, fruit, nut, and field crops, NASS has devoted focused efforts in building a list that had much better coverage of fruit, nut, and vegetable operations.

Recently NASS has worked with ASCS to investigate better use of ASCS records for list building purposes. The NASS tested a procedure that identified approximately 43 percent of ASCS crop records not matching NASS records as unlikely farm operations. In the past a large proportion of the ASCS nonmatch records have been out-of-business or deceased farm operators or landlords. Without a method to identify the likely nonfarm records, NASS use of ASCS data files has resulted in large numbers of records whose farm status needed to be resolved with a resulting low yield of farm records. Additionally, NASS has arranged with ASCS to have direct access to the ASCS records using a relational database. This will enable NASS to extract only those records with specified characteristics and only the items on each record that are of use to NASS. Computer processing costs are expected to decrease, and list building procedures are anticipated to be more effective at increasing the coverage and quality of the NASS list frame.

The NASS plans to provide each state with annual listings of ASCS large or specialty operations that do not match NASS records. Operations producing commodities with insufficient present records for targeted sample designs (such as those new in the estimating program) would be selected. General matches of the entire state ASCS data files to the NASS list would be conducted only once every three years, rotating states each year. The screening procedure would be used to identify the nonmatching records that would be extracted. This approach should result in a more effective use of NASS and ASCS resources to produce higher quality NASS list information. This is an important consideration as determination of actual farm status of potential list records is a time consuming and resource intensive effort.

3.3 List Frame Farm Operator Composition

Although both organizations build their lists by acquiring lists of addresses associated with agriculture, this does not ensure that each record represents a farm operation. For both organizations, the identifier and agricultural data available for each list address does not generally provide adequate information to determine whether or not an address represents a farm operation. To use the list frame effectively for either a census or survey, farm status needs to be identified for each list record. Mail or telephone surveys, personal enumeration or follow-up, or statistical modeling have been used to accomplish this objective with input records. The accuracy of these procedures affects the

size of the census mailing and the efficiency of the NASS sample designs.

Prior to both the 1978 and 1982 censuses, the Census Bureau mailed the Farm and Ranch Identification Survey to approximately 3.0 million addresses that had questionable farm status or were potentially duplicate addresses. The report form contained a set of initial questions which, if answered as "no", allowed addressees to skip the remaining questions. The final 1982 census mailing consisted of 3.65 million addresses of which 1.2 million were respondents to the identification survey classified as representing potential farms and .5 million were survey nonrespondents.

In both the 1987 and 1992 censuses, the Census Bureau used statistical classification analysis to aid in the identification and removal of nonfarm addresses from the list. Census classification analysis is a nonparametric method where previous census record characteristics such as the source of the mail list address, number of source lists on which the address appeared, expected value of agricultural sales, geographic location, and past census farm status were used to separate records into groups according to the proportion of expected census farms and build prediction models. The models were then applied to the 1992 list records and provided an estimate of the probability that a current mail list addressee with that group's characteristics operated a farm (Owens, 1989). The groups of addresses least likely to represent farms (farm probability less than 18 percent for the 1992 census) were removed from the mail list.

An evaluation of the 1987 classification tree analysis (Schmehl, 1990), included a sample survey of records dropped from the census list, belonging to model groups whose proportion of farms was expected to be 11.7 percent or less. Approximately 14.6 percent of these records represented farm operations (46.4 percent survey response), or approximately 25,500 of the 175,000 records dropped represented farm operations. From a separate coverage evaluation program, an estimated 242,850 farms were not on the final census list. Using the two estimates, about 10 percent of the farms not on the mail list (25,500 of the 242,850) were on the preliminary list and the remaining 90 percent were either on the list of nonfarm addresses which were dropped or not on the list at all. This evaluation and others indicated that the analysis accomplished the objective of identifying the groups of records with questionable farm status and provided a reasonably good estimate of the impact of reducing the size of the final mail list on census coverage. For the 1992 census, refinements were made to the model, including the use of standardized computer software and the production of unique state (rather than regional) model groups.

Most NASS State Offices use an annual mail or telephone criteria survey to collect data on operation identification and commodity production. The universe, frequency, and timing of this survey varies considerably from state to state. Field work conducted by personal enumerators often supplement or reconcile the mail and telephone data collection. This survey is used for addressees gleaned from lists acquired by the office and for active farm records that do not have current survey information. New farm records are placed on the active state sampling frame with their agricultural (control) data, and information and data are update for existing records. Nonfarm records are retained on the inactive list frame with a status code that distinguishes records of deceased, retired, or out-of-business operators, of landlords, and of those without agricultural activity.

Annually NASS estimates the number of active list records that do not currently represent farms. This estimate is the difference between the number of active farm records and the sample expansion of the enumerated units in the June Agricultural Survey area sample that are also on the list. In 1992 this estimate was 411,924 or 26 percent of the farm records compared with 474,446 or 28 percent in 1989. Additionally, the NASS measures the percent of sampled addresses in the Quarterly Agricultural Survey that were identified as nonfarms - 10.5 percent in 1992 as contrasted with 13 percent in 1989. This is an estimate of nonfarms on the active farm list frame that were identified as having the commodities of interest.

4. ACCURACY OF RECORDS

The accuracy of both the address and data contained on each list record is critical for a quality census or survey data collection. The accuracy of the address information affects the ability of the survey organization to locate the addressee. The accuracy of the agricultural or control data influences the efficiency of the Census and NASS survey designs. If cases are included in samples inappropriately, the cost and quality of the data collection is compromised.

4.1 Procedures Affecting Accuracy

Specific procedures are used in census mail list development to increase the accuracy of record information (Gaulden, 1990). The name control standardizes the format for each name on the list for comparisons and is essential for identifying potential duplicates during the initial linkage on identification number. The processing code facilitates the use of the

most reliable information (both address and data) in the final record. Both source and size codes are important for sample stratification, classification analysis, census processing, and for evaluating the census mail list.

After every survey, NASS state offices make on-line name and address changes and update the record status code used for effective sampling and data collection. Active record status codes identify refusals, records for special handling, operators linked to additional operations, etc. Inactive record status codes identify deceased, retired, or out-of-business operators, landlords, partner of primary operator (with linkage specified), multiple operations, etc. Inactive source codes such as ASCS, list frame, criteria survey, and other data surveys are retained with the agriculture data. The NASS survey data is captured into a file that is held until the annual classification period. At that time all data for a given record is compiled and the "best" value (generally the largest current year value) is selected by a ranking process as control data for the record. States often specifically extract records without control data for major data items for inclusion in criteria surveys.

4.2 Measures of Accuracy

Several measures from the census data collection permit an assessment of the accuracy of address information. One such measure is the number of forms that are undeliverable as addressed (UAAs) - addressee is deceased, name or address has changed, or another situation is indicated. In 1987 there were 148,252 UAAs (3.6 percent of the census mailout). This compared with a lesser number of 82,792 UAAs (2.3 percent) in 1982 where, 446,000 UAAs had been previously removed from the preliminary 1982 list using information derived from the Farm and Ranch Identification Survey.

Another indication of the accuracy of 1987 census address information was obtained from results from an intensive follow-up to the Nonresponse Survey -- a survey conducted prior to completion of the census data collection for each state to estimate the percent of farms among the nonrespondents. The follow-up examined a sample of 1,263 nonrespondents to the survey. In early 1989, certified mailings, telephone and personal enumeration contacts were conducted on these cases to obtain further information about the nonrespondents. A total of 30.2 percent of the sample cases were UAAs. Of the 31.2 percent for which personal follow-up was attempted, over half had address changes before or during the census data collection. Either the corrected address information had not been received from the Post Office as requested or the Census Bureau had not successfully incorporated the address change into its mail list.

The expected sales code is an extremely important data item on both the census and NASS lists as it is used for sample stratification. Tabulations of expected sales code on the census address record by actual total value of agricultural products sold (TVP) in the census provides an indication of the accuracy of this code. In 1987, approximately 34 percent of all list records had the same coded value for TVP on the census as on the mail list; 72 percent had a value within one size code in either direction.

The NASS computes the correlations of control data (list frame data) with actual survey values for all data items used for sample stratification. In 1992, the U.S. correlations ranged from .766 to .917 for these items with hog, cattle, and land in farms data more closely correlated than cropland and storage capacity. The NASS also measured the percentage of these data items updated in each of the previous six years. Between 51.2 and 69.3 percent of these data items on records selected for at least one survey have been updated in the past two years. States are provided with estimates of percent of the records sampled on the active frame for which these data items are more than five years old. State offices are advised to review the timing of list frame updating and their data selecting and capturing procedures if they have correlations below .40 or if their state has a high percentage of data for any of these items that is more than five years old.

5. DUPLICATION OF LIST OPERATIONS

In the census, list duplication leads to the potential for duplicate census enumerations. For the NASS, list duplication results in incorrect sample weights. For both agencies, the primary objective of list computer and clerical linkage rules is to increase the accuracy of the matching procedure. However, there is a delicate balance between eliminating list duplication and maintaining list coverage. The census computer and clerical procedures have been designed to identify almost exact matches and have relied to a large degree on self-identification of duplicate census report forms, thus increasing coverage with the accompanying risk of increased duplication. The NASS assumptions during list building focused on reducing duplication; the list resolution process for updating and maintenance focuses on increasing coverage. Removing list duplication presents operational challenges for both organizations.

Recently, both agencies have faced an additional challenge to have current and nonduplicated addresses on the list frame. Rural route and box addresses are being changed to street addresses to

facilitate finding addresses for emergencies. If no procedure is instituted in a computer matching process to detect this situation, multiple records for an addressee may be retained on the list frame.

5.1 Procedures Affecting Duplication

In the agricultural universe, farm or ranch operators often have multiple operations with an identified operator participating in one or more partnerships as well as an individual operation. The NASS uses a cross-referencing system of list frame identification numbers. The census mail list compilation flags historically identified operations with different names as possible partnerships or corporation (PPC records). A PPC record flag is used to prevent automatic computer deletion of records as duplicates, causing all paired addresses to be clerically reviewed. During the preparation of the 1992 census mail list, a telephone enumeration of selected PPC addresses was conducted to determine linkage and reduce duplication on the mail list. Of the 25,000 record sets contacted (a total of 107,820 records), 45,464 records were deleted as duplicates.

Recently the sampling unit on the NASS list frame was changed from farm operations to farm operators in order to target a unique unit. This change was implemented by first matching the NASS list against itself using the Fellegi-Sunter procedure to identify matches and probable matches. The matches were identified as inactive records with linkage to the active record. Each state office clerically reviewed all probable matches, determining the name and address of the operator. Operations with multiple operators had the primary operator identified on the file. Additionally, multiple operations for one operator were identified and were linked with that operator.

The Census Bureau introduced a probability based linkage system (variant of the Fellegi-Sunter procedure) for name and address matching in 1992. A modification of this system was used to match 1990 Post Enumeration Survey records to the 1990 Decennial Census. This system permits the user to specify the degree of certainty (threshold values) desired for matched records. Eight percent more duplicate addresses were identified during the first linkage phase; one percent more were identified in the second phase. The evaluation is not yet complete that will provide the percent of increase in duplicate identification attributable to the new linkage system.

The NASS had developed a list duplication check program matching records on numeric fields such as SSN, EIN, and telephone number whose primary use was to identify duplicates after samples were selected from the list frame. Then the probability of selection of each sampled unit was adjusted for list duplication. This program is now additionally used prior to selecting list records from those classified for particular samples. Those identified potential duplicates that are sampled are then flagged for review during the survey edit process.

The Census Bureau used several new (or previously used) procedures to help identify duplicates during data review and processing of the 1992 census. To facilitate self-identification of duplicate report forms, the Census Bureau provided instructions to the respondent on the form and information sheet. A statement was added to the 1992 census envelope to remind the respondent to return all duplicate report forms in the envelope. A duplicate search operation was conducted during data review in all states, sorting all records alphabetically within counties and matching on telephone number and important data variables. Although telephone numbers were first available in the 1987 keyed data, telephone number matching was only used during processing in a few states.

5.2 Measures of Duplication

The census coverage evaluation program measures error in report form farm classification and in list duplication as well as farms not on the mail list. Nonfarms classified as farms and duplicate operations contribute to overcounted farms in the census. The total number of 1987 estimated overcounted farms (135,600) was very similar to the 1982 estimated number (113,623). However, the proportion of the overcounted farms that represented duplicate operations changed from 17 percent in 1982 (19,062 farms) to 47 percent in 1987 (63,290 farms). This increase in 1987 was primarily attributed to the lack of a precensus screening survey.

6. UNIVERSE COVERAGE OF LIST FRAME

The objective of a quality list frame for censuses or surveys is to provide as complete coverage as possible of the target universe. This is an extremely difficult goal for both NASS and the Census Bureau because of the impossibility of identifying operations in the target universe for inclusion in the list frame. Because the census of agriculture list is the basis for a mail enumeration, the overall coverage of the frame and the accuracy of the record information are extremely important. For the NASS, complete frame coverage of the universe of farm operators is not as important since a land based area frame is used to estimate for list incompleteness. Neither is the accuracy of address

information an overriding consideration for NASS as many contacts are initially made through a personal or telephone enumerator rather than a mail delivery person. Both organizations have formal means to evaluate the coverage of their list frame.

6.1 Procedures Affecting Coverage

The NASS has allocated more resources in recent years to increase the coverage of its list frame. The primary mechanism for this has been to use criteria surveys with ASCS records not matching the NASS list frame. As previously indicated ASCS list size constraints have limited the effectiveness of this approach without the ability to screen ASCS records. The NASS has not been able to experience very large increases in its list coverage with this approach. New procedures (described in Section 3.1) have been developed that hold promise for increases in coverage by more effective use of NASS resources.

The Census Bureau has conducted formal coverage evaluation programs for each census of agriculture since 1945. The program measures the accuracy and completeness of farm counts and selected data items and seeks to identify situations that lead to coverage error and to reveal data deficiencies and problems associated with census processes. The evaluation is conducted using an independent sample selected from the list to measure classification and list (duplication) error and the NASS June Agricultural Survey (JAS) to estimate the number of address records not on the census list.

6.2 Measures of Coverage

The much more rigid 1987 and 1992 census size constraints and restrictions on precensus screening necessitated that the resulting list be smaller, yet have a higher proportion of farm operations to ensure good coverage. Data from the census coverage evaluation programs indicate that coverage of the census farm universe is not complete. However, coverage of agricultural production has historically been above 95 percent for all censuses. Historical coverage estimates show that net farm coverage of actual farms has ranged from 85.0 to 92.8 percent for all of these censuses except 1978 where using both a list and area frame census achieved a coverage of 96.6 percent. This methodology substantially improved the state and U.S. level coverage of the census, particularly for farms with sales of less than $2,500 where the census enumeration is least complete. In 1978, the percent of these farms not included in the census was 6.5 percent compared with 28.6 percent in 1982 and 32.3 percent in 1987. Budget constraints have not permitted the use of dual frame methodology in subsequent censuses.

The coverage evaluation program for the 1987 census of agriculture was enhanced to provide estimates of farms not on the census mail list at the state level as well as more reliable estimates of incorrectly classified operations and duplicate operations on the census mail list. The percent of estimated farms on the census mail list was approximately the same in 1987 as in 1982 -- 89.2 percent contrasted with 89.4 percent. An estimated 98.6 percent of the land in farms, 92.3 percent of the crop farms, and 87 percent of the livestock farms were on the 1987 census list. Regional estimates of percent of census published farms not on the mail list are also produced, indicating that the list provides the most complete coverage for the Midwest region and the least complete for the Northeast region.

The reduction in size of the total mail list and the lack of a screening survey did not adversely affect the coverage of the mail list. The changes in the source lists for the 1987 census, improvements in the quality of the source records, and the effectiveness of the classification analysis contributed to maintaining the previous level of list coverage despite the drastic reduction in total list records included in the census. The estimated number of duplicate operations on the 1987 list substantially increased over 1982 -- 2.8 percent contrasted with .8 percent. No 1987 census procedures proved as effective in removing list duplication as the precensus Farm and Ranch Identification Survey.

The NASS has used the area sample from its JAS to estimate universe values for number of farms (by types of crop and livestock) and land in farms since 1985. From this sample it is possible to estimate the coverage of the NASS list for important data items. Studies in 1992, estimated 56.3 percent of farms, 77.6 percent of land in farms, 57.8 percent of crop farms, 53.6 percent of livestock farms, and 37.6 percent of specialty farms were covered by the active records on the NASS list frame. Estimates of the percent coverage by the NASS list for 1990, 1991, and 1992 demonstrate a gradual increase in coverage for these data items -- coverage of number of farms increased 2.6 percent, land in farms increased 3.1 percent, crop farms increased 6.1 percent, livestock farms increased 2.2 percent and specialty farms increased 5.1 percent. As with the census list, the NASS list coverage varies by region.

7. FRAME COMPILATION COST

The costs of compiling an agricultural list frame include salaries of professional staff who design

and implement procedures; procurement of source records; computer processing, linkage, and geocoding of those records; cost of screening list addresses to determine farm status; and salaries and travel costs for field or clerical staff reviewing address information to determine status and potential duplicates. Deriving separate costs for list frame building and maintenance from other survey costs is difficult.

A large proportion of the cost is associated with salaries for professional staff. At the Census Bureau a high level of work occurs for the development, implementation, and evaluation of the mail list during three years of the census cycle. Planning and research occur in the remaining two years. Professional staff work on the list requires a minimum of three statisticians during the entire cycle and three computer programmers during the three year development, implementation, and evaluation period.

At the NASS, list building and maintenance is an ongoing program. A staff of eight people located in NASS headquarters is responsible for developing computer procedures for matching lists and selecting list samples, for developing procedures for maintaining and updating record information, for evaluating the completeness and effectiveness of the frame, and for assisting state office users through documentation, training and consultation. One to two staff in each of the 45 State Offices have on-going state list building and maintenance responsibilities, with other staff used as needed. Extensive computer support has been required for design of new list frame systems.

The Census Bureau obtains source records for minimal cost as they are output of other statistical and administrative data file preparations rather than separate data collection costs. For example, in 1987 the Census Bureau paid the NASS $30,000 for its 2.4 million records and IRS approximately $125,000 for its 6.0 million records. The NASS receives the ASCS list and most of its commodity and trade association lists at no cost. Any costs are routinely associated with the cost of file preparation. Often the most significant costs associated with using another list source are additional programming resources required to standardize the formats of different lists. As programming resources are scarce at both agencies, additional list sources are added selectively if the list format does not meet agency requirements.

The number of records used in the census mail list linkage or incorporated into NASS list building efforts affects the overall cost of computer linkage and resulting staff costs for clerical review or field follow-up. The number of records was 1.1 million (8 percent) less for the 1992 census than for the 1987 census, and 5.4 million (30 percent) less in 1987 than in 1982. The arrangements that NASS has recently developed with ASCS will reduce the number of records in the selected files to be matched to NASS records and the resulting nonmatches. Both lists are computer processed on mainframe computers with their associated costs and overheads. The Census Bureau estimates of these costs are not tracked separately but included in the overall cost of census processing. The NASS has separate cost estimates for the headquarters and state office mainframe processing. The headquarters costs for list maintenance are approximately $450,000 per year, but the state costs are not easily separable from other survey processing costs.

The computer record linkage rules for both organizations are designed to avoid computer deletion of potential duplicates unless there is a high degree of certainty that the potential duplicates are matches. Less stringent matching rules could decrease the number of potential duplicate sets provided for census clerical review, thus reducing clerical staff costs. An indication of staff costs for this clerical work is provided by the number of potential duplicate sets of census records prepared for review during the first phase of linkage--573,148 in 1992, 767,448 sets in 1987, and 1,332,000 sets in 1982. The Census Bureau and NASS, in its initial list building, controlled the cost of this review by selectively setting the parameters for designating records as potential duplicates. At NASS clerical staff and enumerators are used to follow-up on potential farm operations resulting from nonmatching ASCS or commodity list records using mailed criteria surveys and telephone or personal contact. Costs are substantially lessened when mail or telephone contacts are used or personal follow-up is employed in conjunction with other data collection efforts.

The census classification analysis with its associated computer and professional salary costs was an inexpensive substitute for the much more costly precensus Farm and Ranch Identification Survey. However, the application was only designed to remove nonfarm addresses from the list. It did not accomplish the other two objectives of the screening survey - to obtain more current address information and to identify duplicate operations. The costs of such an independent data collection are relatively high. Although the costs of any survey is affected by the size of the survey mailing (3.0 million in 1982), the marginal cost of additional survey cases is small with a large scale data collection.

8. SUMMARY

Building a high quality list frame for large scale data collection is a difficult task. It requires an

ongoing program of research, evaluation, and development whether or not the list is maintained and updated periodically as the NASS list is or recreated cyclically as the census of agriculture list. Changes in source records and postal delivery procedures affect the list compilation process, requiring new list techniques. The purpose for which the list is intended is an important factor in determining the quality requirements for the list.

Compiling and maintaining a list of agricultural operations to conduct either censuses or surveys has a unique set of challenges. An agricultural list frame is unlikely to ever have a complete list of farm operations due to the high turnover in agricultural operations. Research following the 1982 census determined that only 71 percent of farms in 1978 were farms in 1982. Maintaining a high level of list coverage requires continual attention to improvements. In the U.S., tax records are essential for achieving a high level of list coverage, with this source uniquely providing approximately 14.9 percent of all identified farm operations in 1987. This is an important source for identifying new operations and those that have gone out-of-business. The NASS uses a number of different sources and procedures to try to accomplish the same objective.

The many different arrangements under which farms and ranches operate will invariably affect duplication in the list. As the Census Bureau discovered in 1987, controls to eliminate duplicate enumeration in the overall census processing system were lacking once the precensus screening survey was eliminated. Several procedures were initiated in 1992 to identify duplicate addresses and reports. This emphasizes the importance of continually assessing changes in methodology intended to increase quality of one aspect of the list in relation to the impact on other quality attributes.

In order to either build or maintain a high quality agricultural list frame, continual evaluation and measurement of the frame characteristics discussed in this paper will be needed. The attributes of the list -- its scope, accuracy of record information, duplication of records, universe coverage of list frame, and cost of list frame compilation -- will need to be reviewed in relation to the program objectives that the list frame serves. Although both census and NASS require a high quality agricultural list frame, the differing program objectives of these frames affect the importance of each of these attributes to the overall quality and functionality of the frame. The primary objectives of the census of agriculture list are farm coverage and uniqueness whereas the NASS objective is to obtain a high level of commodity coverage. These objectives affect the assumptions underlying the list frame procedures described in this paper.

REFERENCES

Anderson, Carter D. (1993), "NASS Long Range Plan for Use of ASCS Data, 1993 - 1997", NASS Internal Staff Report.

Arends, William L. (1977), "Methodology for the Development, Management and Use of a General Purpose List Sampling Frame", NASS Internal Staff Report.

Coulter, Richard and James W. Mergerson (1977), "An Application of a Record Linkage Theory in Constructing a List Sampling Frame", *Proceedings of Computer Science and Statistics: Tenth Annual Symposium on the Interface*, Gaithersburg, MD.

Fellegi, Ivan P. and Alan B. Sunter (1969), "A Theory for Record Linkage", *Journal of American Statistical Association*, pp. 1183-1210, December 1969.

Gaulden, Tommy W. (1990), "Development of the 1987 Census of Agriculture Mail List," U.S. Census Bureau internal report.

Geuder, Jeffrey (1992), "1992 NASS List Frame Evaluation", NASS Survey Management Division Report, eighth of a series.

Lynch, Billy T. and William L. Arends (1977), "Selection of a Surname Coding Procedure for the NASS Record Linkage System", NASS Research Division Report.

Owens, Dedrick, Ruth Ann Killion, Magdalena Ramos, Richard Schmehl (1989), "Classification Tree Methodology for Census Mail List Development," *1989 Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Schmehl, Richard, Magdalena Ramos (1990), "Evaluation of Classification Tree Methodology for Census Mail List Development," *1990 Proceedings of the Section on Survey Research Methods,* American Statistical Association.

1987 Census of Agriculture Coverage Evaluation Report, Volume 2, Part 2.

Scope and Methods of the Statistical Reporting Service (1983), NASS Miscellaneous Publication No. 1308, Washington, D.C.