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 ($N_{.1} = 40$ and $N_{.2} = 70$) than the corresponding strata of survey I ($N_{1.} = 30$ and $N_{2.} = 80$). 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, $n_{2.}^{(1)} = 10$, to the population in both stratum 2 of survey I and stratum 1 of survey II.

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

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

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

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

$$\vdots$$

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

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

$$\vdots$$

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

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

where

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

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

$$
\begin{array}{|l}
\hline
N^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{h,\ 1\ ,\ldots,1,1,\ldots,\ 1\ ,\ 1} \\
\hline
N^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{h,\ 1\ ,\ldots,1,1,\ldots,\ 1\ ,\ 2} \\
\hline
\end{array}
$$

$$\vdots$$

$$
\begin{array}{|l}
\hline
N^{(k,k+1,\ldots,K,1,\ldots,k-2,k-1)}_{h,h_{k+1},\ldots,h_K,h_1,\ldots,h_{k-2},h_{k-1}} \\
\hline
\end{array}
$$

$$\vdots$$

$$
\begin{array}{|l}
\hline
N^{(k,\ k+1\ ,\ldots,\ K\ ,\ 1\ ,\ldots,\ k-2\ ,\ k-1\ )}_{h,H^{(k+1)},\ldots,H^{(K)},H^{(1)},\ldots,H^{(k-1)}{}_{-2},H^{(k-1)}{}_{-1}} \\
\hline
N^{(k,\ k+1\ ,\ldots,\ K\ ,\ 1\ ,\ldots,\ k-2\ ,\ k-1\ )}_{h,H^{(k+1)},\ldots,H^{(K)},H^{(1)},\ldots,H^{(k-1)}{}_{-1},H^{(k-1)}} \\
\hline
\end{array}
$$

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

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

(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

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

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*,

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

$$\vdots$$

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

$$\vdots$$

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

where component $k$ of the vector is 1 if the unit is sampled for the $k$th 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 $i$th assignment configuration, represented by the $i$th 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)

$$\begin{aligned} & f(x_1, x_2, \ldots, x_{2^K}) \\ = \; & b_1 x_1 + b_2 x_2 + \ldots + b_{2^K} x_{2^K} \\ = \; & \vec{b}\,\vec{x}^T \end{aligned}$$

subject to the $K+1$ linear constraints

$$\begin{cases} \vec{v}_1 x_1 + \vec{v}_2 x_2 + \cdots + \vec{v}_{2^K} x_{2^K} = \vec{m} = \\ \quad (m^{(1)}, m^{(2)}, \ldots, m^{(K)})' \quad \leftarrow \text{K constraints} \\ x_1 + x_2 + \cdots x_{2^K} = M\,, \end{cases}$$

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

$$\begin{aligned} & \vec{v}_2 x_2 + \vec{v}_3 x_3 + \cdots + \vec{v}_{2^K} x_{2^K} \\ = \; & (m^{(2)}, m^{(3)}, \ldots, m^{(K)})' \end{aligned}$$

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.

# USDA'S ANNUAL FARM COSTS AND RETURNS SURVEY: IMPROVING DATA QUALITY

Bob Milton and Doug Kleweno, Estimates Division, National Agricultural Statistics Service
Rm. 5912-South Building, 14th & Independenc Ave., S.W., Washington, D.C. 20250

KEY WORDS: finance survey, data quality, data sharing, respondent burden

## Survey Description and Use of Data

The Farm Costs and Returns Survey (FCRS) is a comprehensive farm finance survey conducted annually by the National Agricultural Statistics Service (NASS) for the U.S. Department of Agriculture (USDA). In total, some 1,300 data items are collected when all questionnaire versions of the FCRS are considered. Information on crop and livestock production, farm expenses, income, debt, assets, and socio-economic and demographic data are collected.

Information from the survey is the basis for USDA estimates of farm expenditures, income, cash flow, wealth, costs of production, and productivity. The FCRS is an integrated survey that provides information on the farm sector, household, business, and enterprise (for major farm commodities). Information from the survey is provided at the U.S. and regional levels and by type and size of farm. Size of farm is determined by value of annual sales. Much of this information is published routinely by USDA's Economic Research Service in its series Economic Indicators of the Farm Sector and in their Situation and Outlook reports. NASS also publishes detailed expenditure data annually from the FCRS.

The FCRS provides the only annual data set at the U.S. level for farm financial, production, and related information. The FCRS data base is used by ERS in analyzing numerous farm program and policy issues annually for USDA and other policy makers.

## Survey Design

The FCRS is a multiple frame, probability survey of U.S. farms. The sample size over the past 5 years has averaged about 24,000 farms, just over one percent of all farms. A farm is defined as any establishment from which $1,000 or more of agricultural products are sold or could be sold during the year. Types of establishments included in the survey are those listed in the Federal Government Standard Industrial Code (SIC) for agricultural production of crops and livestock - major group codes 01 and 02.

Samples are selected from two sources. The first source is a list of operators of farms and ranches. Control data on type of farm and size are used to stratify the list. The list frame represents the larger, more specialized operations. The second source is an area frame where the continental United States is divided into small area sampling units, each with a known probability of selection. The area frame sample focuses on collecting data on smaller operations, less than $20,000 in annual sales, plus larger operations that are not on the list. Data for the area frame operators not on the list are used to measure the incompleteness of the list.

The survey is designed to provide reliable data at the regional level which represents 10 geographic groups of States with similar production practices. At the U.S. level, the coefficient of variation (C.V.) is about 2-5 percent for major expenditure and income items. C.V.'s at the regional level are generally in the range of 10 to 20 percent. The extent of nonsampling errors is not known. To minimize nonsampling errors, data collection procedures are uniform and consistent across the Nation by using extensive training and field supervision of data collectors.

The FCRS is designed to provide estimates of several types of information. Accordingly, several versions of the FCRS questionnaire are used to collect the types of information. Depending upon the questionnaire version, additional data are collected on cost of production for specific commodities on a 4-5 year rotation, on socio-economic and demographic data, and on detailed expenditure and income data. All questionnaire versions have basic income and expenditure questions so that all questionnaires are additive to generate certain basic financial information. The different questionnaire versions provide additional independent estimates of specific information depending upon questionnaire purpose.

## Survey Problems and Data Quality

The largest obstacle confronting the FCRS evolves around the large amount of detailed data collected from a shrinking population of farmers. Some 1300 separate data items are collected annually on the FCRS. Many of the these items are related to the costs of production surveys where minute detail is needed in constructing costs of production budgets and models.

The more detail collected, the greater the respondent burden becomes. Average interview time for the 1990 survey was nearly 1 1/2 hours overall (Rutz and Cadwallader, 1991). The average interview time for the 1990 cow-calf costs of production questionnaire version was nearly 2 hours and interviews of 3-4 hours were common (Appendix Table 1).

The interview time requirements for the FCRS is a major reason the survey response rate is relatively low (10-20 percentage points lower) compared with other NASS surveys and continues to erode (Appendix Maps 1 and 2) (Rutz and Cadwallader, 1991). Over the past five years (1987-92) the response rate for the FCRS has fallen from 73 to 66 percent. In research conducted on reasons for nonresponse to the 1990 survey, one-fourth of all refusals indicated they would not take time to complete the survey (Appendix Table 2) (O'Connor 1992). The overall refusal rate for the 1991 survey was 25 percent, but was as high as 33 percent for the corn costs of production questionnaire version. In two States, the overall refusal rate was above 50 percent. The response rate is also lower among the large farms. The response rate for the largest farms sampled from the list frame, farms with annual sales over $500,000, for 1990 was 57 percent compared with 69 percent for all farms (Appendix Table 3) (Rutz and Cadwallader, 1991). Field offices have also indicated that large farms have a greater tendency to refuse in the future once having completed a lengthy interview.

The higher level of nonresponse for the large farms is particularly critical with regard to data adjustment for nonresponse. Data are adjusted for nonresponse at the strata level within State by the ratio of good responses plus inaccessible and refusal samples to good responses. In many cases this adjustment more than doubles the expansion factor for responses from the largest farms, annual sales of over $500,000. This strata of farms accounts for only two percent of all farm numbers but over two-fifths of total farm expenditures and gross income.

Beginning with resummarization of the 1991 data, the nonresponse adjustment was modified so that all refusal and inaccessible samples were assumed to have positive farm data (Turner, 1992). Field enumerators were instructed to verify that refusal and inaccessible samples had positive farm data, some type of crop or livestock production. The modified adjustment removed the count of operations without positive farm data, out of scope operations, from both the numerator and denominator. The resulting larger nonresponse adjustment factor increased the expansion of total U.S. expenditures and income by about 9 percent. The increase due to the change in the nonresponse adjustment was greater than what was assumed before research proved otherwise. The greatest increase occurred in the upper strata, or large farm classes, where it had been assumed that there were fewer screenouts or out-of-business operations.

The nature of the FCRS, to collect personal financial data, is another major contributing factor to the relatively lower response rate on the FCRS. Beyond no reason given, the nonresponse research indicated that the second most frequent reason for refusing to complete the survey questionnaire was that the information was too personal. Besides the 25 percent that refused the initial interview, refusals or "don't knows" to some questions accounted for as much as 15-16 percent of expanded data for some items, specifically value of farm assets and landlords' share of government payments (Appendix Table 4) (Morehart and Johnson, 1992). On average, expanded data for refusal items amounted to 1-2 percent. For refusal or "don't know" items, data are imputed by combining all U.S. data into one file and calculating average by type and size of farm for the missing items. This level of imputation occurs after the raw survey data are considered "clean".

A thorough clerical and machine edit is also run on the raw data as it is received, prior to the imputation edit. Research on this edit concluded that the edit has little effect on the final results and that the small effects are accounted for by a very few reports (Hoge and Willimack, 1991). Nearly half of the edits move respondent data to the proper cell with little or not effect on data expansions. The same is true for detailed editing for incomplete allocation of aggregate reported data.

Once the machine edit is completed, the data are summarized and an outlier review takes place. An outlier is defined as a report whose expanded data account for 5 percent or more of the regional total or one-half of one percent of the U.S. total for major data items (Statistical Methods Branch, 1992). The outlier adjustment process moves the report to the largest operator stratum where all large operations have the same expansion factor. If it is further determined by the outlier review board that the extreme operation is unique in itself or is similar to only a few operations, the expansion factor is further reduced.

One additional adjustment is made to the FCRS data to ensure complete farm coverage. Data are adjusted by sales class at the regional level by the ratio of FCRS expanded number of farms to estimated USDA number of farms (Statistical Methods Branch, 1992). The area frame expansion for the FCRS has been historically based upon a sample of resident farm operators. This expansion of farm numbers is generally about 15 percent below the official estimates. The reason for the incompleteness of the farm coverage from the area frame is largely due to the inability to pick up farm operators, especially small operators, in the urban and suburban land units, segments. Adjustment for undercoverage of farms was initiated with the resummarization of the 1991 data and added about 3 percent to total expenditure expansions.

## Future Direction on the FCRS

Future direction on the FCRS should focus on increasing response rates. Of utmost importance to increasing response rates is reducing interview time. Preliminary plans are to expand the use of the aggregate expenditure questionnaire version that eliminates the detail or breakout of component expenditures from the group total and collects no commodity costs of production data. The interview length was reduced by about one-fourth hour for the aggregate questionnaire compared with the detailed expenditure questionnaire during 1992 tests. Expenditure data for the farm operation that is part of the cost of production questionnaire version will also be collected at only the aggregate level.

Expanded use of a global short version expenditure questionnaire to the operational level also fits within future plans. This questionnaire of 16 pages is even more abbreviated in length than the aggregate version. A global short version questionnaire was tested in the farm finance follow-up to USDA's Chemical Use Survey in 1992. Preliminary review of data from this global short version is promising with regard to collecting data at the aggregate rather than component level. Response rates for the global version were significantly higher than for the FCRS in the two States conducting the farm finance follow-up survey. In Louisiana, the response to the global short version was 76 percent compared with 65 percent for all versions of the FCRS and in Minnesota the response was 62 percent compared with 56 percent for the FCRS. Much of this increase in response rates is however attributable to the screening out of refusal and out-of-business operations before arriving at the sample size for the farm finance survey.

The high level of respondent burden on the larger farms due to frequent contacts for a variety of surveys causes a need to concentrate on sampling schemes that will reduce the number of contacts. Sampling plans are being considered that integrate the needs of several surveys with one sample selection using basically nonreplacement sampling of strata that meet the needs of all the surveys. Preliminary post-survey research analysis covering four major surveys in three States during 1991-92 indicates a potential reduction in individual respondent burden, or number of multiple contacts, of 60 percent (Preliminary research by NASS researchers Dr. Charles Perry and Jim Burt).

Current list building activities should enhance the sampling work. List building activities this year concentrated on trying to improve coverage on farms with annual sales of $100,000 or more. In 1992, list coverage at the U.S. level for farms with annual sales of over $100,000 was 89.3 percent (Geuder, 1992). The goal for 1993 is to improve the coverage of these larger farms to 95 percent. List concentration on adding large farms and improving their control data should enhance sampling and improve data accuracy due to better overall stratification and coverage.

Farm coverage for the FCRS should improve for the 1993 survey due to the switch to a weighted area estimator. Since all area tracts (separate operations within the land segment) and not just resident operator tracts will be eligible for selection, the sampling universe will be larger, reducing respondent burden for resident farm operators and possibly improving response rates. Data for selected area tracts will be expanded based upon the ratio of land

within the tract to land in the entire operation. This weighted estimator reduces the undercoverage bias due to missed area frame farms, especially farm operators living near or in urban and suburban areas, because data are associated with the location of the farm rather than the location of the operator's residence. Starting with the 1993 survey, the current procedure of adjusting data for farm coverage by the ratio of estimated number of farms by sales class to survey expanded number will be reevaluated.

An important factor in improving response rates that needs more consideration is the perception of the survey by the field enumerators conducting the face-to-face interviews. Enumerators play an important, if not the most important, role in obtaining survey response. Most respondent decisions to participate are heuristically based (Groves, Cialdini, and Couper, 1992). Enumerator experiences and expectations affect their ability and motivation to maintain interaction with the respondent. If the FCRS is presumed to be too much of a respondent burden, the questions too personal or too difficult in nature, and data of marginal value to users, response to the survey will suffer (Allen 1993). This situation can be addressed by getting the questionnaire length to a manageable level, providing additional training to enumerators, and "selling" the survey to the enumerators and public.

A task group has been formed within NASS to investigate the low response rate on the FCRS. The task group believes that field enumerators need additional, more specific, training to better handle potential refusal and inaccessible (by respondent choice) respondents. Role playing and special case situations need to be a basic part of training. Enumerators need more training on interviewing techniques, scheduling, and on the purpose and need for the survey.

Above all, field enumerators need to be convinced of the importance of the survey in order to "sell" it. NASS management in Headquarters and the States need to make additional efforts to demonstrate the importance of the FCRS to enumerators. This starts with more public relations work on the FCRS. Studies have shown that public relations more focused to gain the support of groups identified with and respected by the target population are helpful (Slocum, Emply, and Swanson, 1956). Historically, FCRS response rates for sugarbeet growers have been higher than other commodity groups because the industry visibly endorsed and encouraged cooperation. States need to work more with the industries, producers, and media throughout the year on the importance of the FCRS.

NASS is also researching incentives as inducement to improve response rates. Pocket calculators were given out on a trial basis to a portion of the FCRS sample in four States during the 1992 survey. An evaluation of the incentive research has not been completed to date; however, initial results suggest some improvement in response rates. Concerns over the effects on participation in other voluntary surveys have been raised.

## Data Sharing

Another issue that is related to survey response is the confidentiality of the survey data relative to its use. Recently, NASS received a ruling from USDA's Office of General Counsel (OGC) on interpretation of the statutes governing sharing of individual record data such as that provided by the FCRS. OGC's interpretation of the statutes allows data sharing to other agencies, universities, and private entities as long as it enhances the mission of USDA and is through a contract, cooperative agreement, cost reimbursement agreement, or Memorandum of Understanding. Such entities or individuals receiving the data are also bound by the statutes restricting unlawful use and disclosure of the data.

It will be NASS policy that data sharing will occur on a case by case basis as needed to address an approved, specified USDA or public need. NASS and ERS have the responsibility to assure data providers that use of the data will be for public good only. NASS will explore opportunities to broaden the use of cooperative agreements with universities and other government agencies. Access to each data set provided to the cooperative party will need to be properly certified as to the confidential aspects of that data set and regulations. Data sets shared by NASS will be used on-site in USDA facilities and will also be returned or destroyed after meeting the specified need. To improve data access, NASS plans to make the FCRS data available to qualifying entities at two of its State offices on a trial basis in 1993.

## Summary

The FCRS is a probability farm finance survey that produces the only annual comprehensive U.S. data set available that combines farm financial,

production, and related information. The survey is the basis for USDA estimates for farm expenditures, income, cash flow, costs of production, and productivity. The detailed and personal nature of the survey is the major reason for the relatively low response rate.

During the past year, data adjustments for nonresponse and undercoverage have been modified to improve quality of expanded data. Nonresponse and respondent burden problems are more concentrated among the large farms who account for the majority of expanded data. In order to improve response rates, future efforts will focus on sampling schemes that reduce the reporting burden on large farms, shortening the length of the questionnaire to lessen respondent burden, providing more training to field enumerators in handling reluctant respondents, and publicizing the survey more to gain public acceptance. In order to improve access to the FCRS data set, NASS will make the data available to qualifying entities at two State office sites during 1993 on a trial basis.

## REFERENCES


1. Allen J. Donald. (1993). "The Interviewer and the Interviewing Process", NASS Staff Report SMB-93-02.

2. Geuder, Jeff. (1992). "1992 NASS List Frame Evaluation", NASS Staff Report SSB-92-02.

3. Groves, R. M., Cialdini, R. B., and Couper, M. P. (1992). "Understanding the Decision to Participate in a Survey", Public Opinion Quarterly, 56:475-495.

4. Hoge, Stanley J. and Willimack, Diane K. (1991). "Analysis of Item Nonresponse, Imputation and Editing in the 1989 Farm Costs and Returns Survey for Iowa and North Carolina", NASS Staff Report SRB-91-09.

5. Morehart, Mitchell J., Johnson, James D., and Banker, David E. (1992). "Financial Performance of U.S. Farm Businesses, 1987-90", ERS Agricultural Economic Report No. 661.

6. O'Connor, Terry P. (1992). "Identifying and Classifying Reasons for Nonresponse on the 1991 Farm Costs and Returns Survey", NASS Staff Report SRB-92-10.

7. Rutz, Jack L. and Cadwallader, Chris L. (1991). "1990 Farm Costs and Returns Survey, Survey Administration Analysis", NASS Staff Report SMB-91-04.

8. Slocum, W. L., Emply, L. T., and Swanson, H. S. (1956). "Increasing Response to Questionnaires and Structured Interviews", American Sociological Review 21:221-225.

9. Statistical Methods Branch. (1992). "1992 Farm Costs and Returns Survey Specifications".

10. Turner, Kay. (1992). "Modification of FCRS Nonresponse Adjustment Procedures", NASS Staff Report SRB-92-08.

# REMOTE SENSING PROGRAM OF THE NATIONAL AGRICULTURAL STATISTICS SERVICE: FROM A MANAGEMENT PERSPECTIVE

BY

GEORGE A. HANUSCHAK

MICHAEL E. CRAIG

NATIONAL AGRICULTURAL STATISTICS SERVICE

U.S. DEPARTMENT OF AGRICULTURE

## I. SUMMARY

The National Agricultural Statistics Service of the United States Department of Agriculture has been utilizing digital earth resource observation satellite data since the launch of Landsat 1 in 1972. There are currently three applied research efforts in the U.S. agricultural statistics program. These are crop area estimation, crop condition assessment and geographic information system (GIS) utilization for farm chemical and other agricultural survey data. These three research applications are in various stages of development and implementation.

The major research application is the use of Landsat thematic mapper data in combination with area sample frame based ground-gathered data to improve the precision of rice and cotton acreage estimates in the Mississippi Delta region of the U.S. Landsat thematic mapper is a sensor on polar orbiting earth resource observation satellites. The crop area estimates are calculated in an operational timeframe and provided to the Agency's Agricultural Statistics Board as input to the official estimates released by the Agency during the crop season. The well documented agression estimator approach is used. A contributed paper at this conference authored by Graham discusses the statistical procedures in detail. The Delta region was selected because of the excellent separation characteristics of rice and cotton from competing spectral land covers and because of the North-South orientation and relatively small growing region compared to the Midwest or Great Plains regions of the U.S. Landsat data is used and regional, State and county level estimates are calculated. In addition county level classification color coded theme map products are provided to the state offices. This project began in 1991 and will be done on an annual basis. The Agency has a long history of similar projects with Landsat Multi-Spectral Scanner Data from 1972-1990. Accurate cost estimates have been kept for the time series 1972-1992 for these projects for cost benefit analysis comparing the new method to the conventional area frame ground-gathered data approach. The statistical measure of performance used is the relative efficiency which is the ratio of the variance of the ground

data only direct expansion estimator (numerator) and the variance of the regression estimator (denominator).

Larger values of the relative efficiency reflect a larger gain due to adding Landsat data into the estimator process. For the 1991 and 1992 Delta project, the average statistical relative efficiency for rice was 3.5 and for cotton it was 3.9. That is, the sample size on the ground would have to be increased by a factor of 3.5 to 3.9 to match the precision of the Landsat-based acreage estimate. These were cost effective improvements in the precision of the State level crop acreage estimates with no additional respondent burden on farm operators.

In addition, county level estimates and crop specific classification (color theme) maps are provided. Statistical methodology for the county (small area) estimator is provided in detail in a contributed paper by Bellow at this conference. The color coded theme maps provide the complete spatial distribution of crops that conventional sample ground gathered data cannot provide.

The second research utilization of complete spatial and remotely sensed data involves the use of vegetative indices calculated from National Oceanic and Atmospheric Administration's Advanced Very High Resolution Radiometer (AVHRR) sensor. The AVHRR is a sensor on polar orbiting weather satellites. NASS has been slow to get into this area because of its very extensive ground-gathered objective yield forecasting and estimation program already provided excellent information on crop conditions and yields. However, due to the daily satellite passes and the spatial nature of the AVHRR data, there is now interest in calculating and mapping vegetative indices similar to the operational program that Statistics Canada has had since 1988. NASS is currently populating a historic data base of AVHRR data and testing the hardware system to support this type of activity. NASS is using the Land Analysis System (LAS) software from the Earth Resource Observation Satellite (EROS) data center in Sioux Falls, South Dakota and NASA's Goddard Space Flight Center. This program is still in the development stage but the goal is an operational program that would sell on a subscription basis special crop condition assessment data and color map products similar to the Statistics Canada program.

The third and newest area is the use of geographical information systems for providing management additional information about agricultural survey data by taking advantage of the spatial aspects of the data and by overlaying several layers of data such as farm chemical applications, soil types, slope and water flow, crop and land use covers, etc. NASS is in the very early stages of the utilization of GIS based data and related analysis. NASS has procured the ARCINFO GIS software system and also has a Sybase relational data base software system that integrates with ARCINFO. NASS is in the process of populating a farm chemical data base and GIS sample survey data layer at the moment. After completion of these tasks, other layers will be considered as analysis goals and potential become better clarified.

Overall NASS is a fairly extensive user of space based remotely sensed data and related spinoff technologies such as the process of electronic digitization of frame and sample boundaries in its U.S. Agricultural Statistics Program. However, in relation to the overall NASS mission of providing agricultural statistics on hundreds of items throughout a year, the portion of NASS's

program that utilizes remotely sensed data is not large. The Agency has, however, been able to successfully supplement its existing probability based (area, list and multiple frame sampling) estimation program by utilizing digital and image space based remotely sensed data for selected geographic areas.

## II. CROP AREA ESTIMATION IN THE MISSISSIPPI DELTA REGION OF THE UNITED STATES

NASS staff used Landsat Thematic Mapper Data to operationally calculate improved crop acreage estimates for rice and cotton in the Mississippi River Delta Region in 1991 and 1992. The Landsat Thematic Mapper was used in conjunction with area frame based ground-gathered data in the form of a regression estimator. The ratio of the variances, also called the relative efficiency, of the regression estimator and the ground data only direct expansion estimator is the measure of statistical gain from using Landsat Thematic Mapper Data.

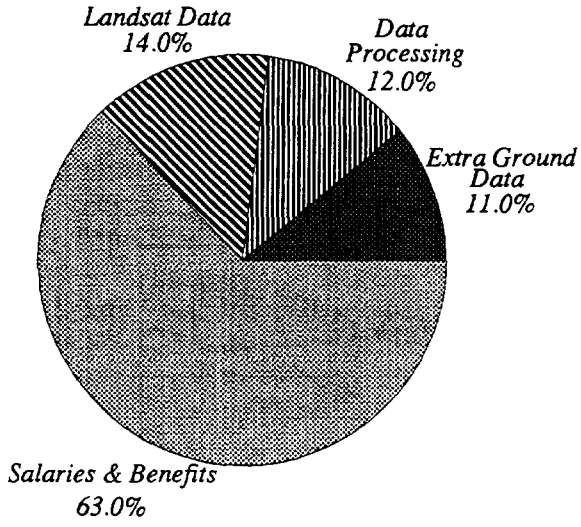
In 1991 and 1992, for rice the relative efficiency averaged 3.5, for cotton it was 3.9 and for soybeans it was 1.9. The relative efficiency can also be interpreted as the factor by which the ground data area frame sample size would have to be increased to match the results of the regression estimator. Due to cloud cover and scene availability factors, the Landsat coverage area was divided into both multitemporal and unitemporal analysis regions. In addition, county level estimates were calculated. Coefficients of variation for the county level estimates for the major rice counties ranged from 3.9 to 10.0 percent. Also, color coded crop classification maps were provided to the State Statistical Offices. The full details of this project are in a recent paper by Bellow and Graham (Aug. 1992).

All the estimates were calculated using the extensive PEDITOR in-house software system developed by NASS and the National Aeronautics and Space Administration Ames Research Center Staff over the years. The PEDITOR system is a quite extensive analysis system for using remotely sensed data in combination with an area frame sample of ground gathered data to calculate regression estimator based crop area estimates and their associated variance. The system has over 100,000 lines of PASCAL code and is used in several other countries around the world. A paper by Jacques Stakenborg (1989) reveals why the European Community's Joint Research Centre chose it over commercial systems for an extensive remote sensing for agriculture statistics project over a ten year period in Western Europe. The main reason PEDITOR was chosen by the European project staff was its efficiency in calculating regression estimates over large land areas. The mosaicing and statistical features are optimized for use with a regression estimator approach. The PEDITOR system's current status has recently been summarized in a paper by Ozga, Mason and Craig (Aug. 1992).

Accurate cost data has been collected and preserved in a data base by project managers since the mid-late 1970's. Thus, NASS has been able to look at Landsat projects (both Multi-Spectral Scanner and Thematic Mapper) from a rudimentary cost/benefit perspective over the years (1975- 1992). The cost side of the equation has been relatively easy to measure. However, as in any cost/benefit analysis the assumptions made about the benefits are a key ingredient to the validity of the analysis. Current total Landsat project costs per State are approximately $175,000. Of

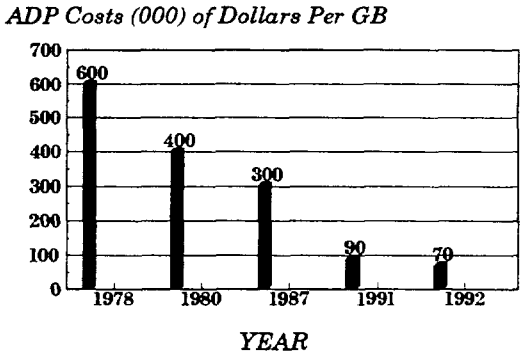
the total, 63% is for salaries and benefits, 14% is Landsat data purchases, 12% is all data processing costs including amortized equipment costs on an annual basis, and 11% is a second visit to ground data sites where fields were not already planted on the first visit (See Figure 1.) In addition, costs per State for the already operational ground survey are approximately $60,000 for the States involved in the project. Landsat project costs have been dropping due mainly to advances in computer technology and in concert with dropping prices for any given level of technology. Ground data collection costs on the other hand are increasing due to inflation in salary, hotel and mileage costs for survey interviewers.

***FIGURE 1: Delta Project Costs***




When total project costs are compared over time and divided by billions of bytes of input Landsat data processed, the project cost drop is dramatic (see Figure 2.) This was due to two main factors. The first has already been cited as the dropping prices of an ever improving computer technology. The second is staff productivity as more States and land areas were done with a constant number of research staff. With the Landsat Thematic Mapper sensor, it is estimated that a relative efficiency in the 3.0-4.0 range is required to be cost effective. Thus, the results for rice and cotton are judged to be cost effective improvements. This is especially the case since the fairly dramatic improvements in the precision of the State level crop acreage estimates do not add to total respondent burden which is a major concern in U.S. agricultural surveys. The county level estimates and maps with measurable precision are additional benefits. However, the success across years and seasons is still dependent on the degree of cloud cover during the critical crop discrimination windows which are usually only 30 - 40 day windows at best. The probability of success for these projects would be increased substantially by having eight day coverage (two Landsat TM systems) instead of the planned one at a time Landsat 6 and 7 systems.

***FIGURE 2: ADP Costs Per Gigabyte (GB) of Import Data (1987-1992)***




## III. VEGETATIVE INDICES

NASS has recently (last 18 months) begun to explore the possibilities of crop condition assessment utilizing vegetative indices from NOAA's AVHRR data. The Agency has been slow to get into this area because of its very

extensive and sound ground-gathered survey data program to forecast and estimate crop yields. The conventional program utilizes both objective crop counts and measurements such as corn ears, ear length and circumference, field and laboratory weights etc. for each crop plus farmer reported yields. Both types of data have long well established time series and provide a relatively high performing system for forecasting and estimating crop yields. The most comprehensive document of the U.S. system for forecasting and estimating yields was by Huddleston (August 1978). For a current update, the Agency survey manuals and a paper at this conference by Birkett would be the best source.

However, due to the daily satellite passes, the spatial nature of AVHRR and complete national coverage, NASS research staff saw some new potential. In addition, close cooperation with Statistics Canada's Agriculture Division enabled NASS to observe their AVHRR vegetative index program which became operational in 1988. These facts combined inspired NASS research staff to initiate a program. NASS has begun a cooperative agreement with the Remote Sensing Laboratory of USDA's Agricultural Research Service to investigate vegetative indices as related to crop conditions. Condition assessment encompasses such topics as comparison of current year crop(s) growth to previous year(s), comparison of crop growth within a given year between States or counties, and drought and crop disease monitoring. The AVHRR-based Normalized Difference Vegetative Index (NDVI) produced biweekly by the EROS data center will be specifically evaluated. Early research in crop condition assessment will center on the evaluation of NDVI color line printer plots, building a historic data base of NDVI and on the potential use of the NDVI for yield models. The AVHRR vegetative index data provides virtually complete national spatial coverage every two weeks. The spatial resolution of the data is one square kilometer. Thus, when combined with other geographic boundaries such as State and county in a GIS, many different geographic levels of data aggregation and comparisons are made possible. Tabular and color theme map data, when put in a GIS, can be aggregated or displayed by any polygon of interest. Thus, the vegetative index data has potential to be one input variable in crop yield models. The Agency plans to value add to this data by using other existing Agency data sets including area sampling frame strata. The subject of the Agency's area sampling frame is covered in detail in an invited paper by Bush and House at the conference. The Agency's area sampling frame and Landsat crop specific classifications can be used as masks to narrow down the polygons of interest for the AVHRR vegetative index. The polygons of interest can then exclude non-agricultural land and in some cases provide crop specific polygons for input to crop specific yield models. A DEC VAXStation workstation has been purchased for this project; it will utilize a current version of the Land Analysis System LAS software developed by the U.S. Geological Survey and the National Aeronautics and Space Administration's Goddard Space Flight Center staff.

## IV. ENVIRONMENTAL DATA AND GEOGRAPHIC INFORMATION SYSTEMS

The newest major addition to the Agency's survey program are farm chemical application data in various forms. Survey programs have been designed and implemented (1989 - current) to measure farm chemical applications at the farm level and at the individual field level on a sample survey basis. As part of the U.S. President's

Water Quality and Food Safety Initiatives, NASS has become the surveyor of farm applied chemicals. As part of these initiatives, the tasks of putting these data in a data base and into a geographic information system were also assigned to NASS. NASS has utilized SYBASE (a UNIX based relational data base system) and ARCINFO (a GIS system) as the software to provide the necessary platforms for storing, retrieving and analyzing the sample survey farm chemical data. Data at the published level and micro data will soon be entered into these systems. Confidentiality of farmer reported data will be strictly protected as only use for official government statistical purposes will be allowed and individual data will not be revealed in any form of publication. In addition, a small pilot project was initiated to look at Global Positioning Systems (GPS) recorders for getting accurate coordinates of field locations. A recorder was used to label points within several sample segments in Ohio. This technology, as reported by many other applications scientists, seems to meet most accuracy needs. However, the up front capital investment in equipment, software, training, etc. was judged to be too high for current Agency applications. However, as costs continue to drop, the GPS technology holds substantial promise for several Agency applications such as GIS, area frames, etc.

## V. IN HOUSE COMPUTER SYSTEMS

To service these requirements, a wide range of microcomputer technologies are interfaced in- house. Large volume remote sensing analyses are performed on a VAXCLUSTER of a MicroVAX 3500 and a VAXStation 3100. Other technology research applications, such as GIS, are performed on a UNIX system which utilizes a SUN 4/380 server with SPARC and SUN IPC workstations (both stand alone and client server forms). Both servers have a 9-track tape and Exabyte tape cartridge capabilities in addition to several disk drives and other peripherals.

Smaller volume analyses utilize 386 and 486 personal computers as stand alone and/or client workstations to both the SUN and VAX servers. All servers, workstations and personal computers are connected together on an ETHERNET network using Network File Server, DECNET and TCP/IP protocols. Peripheral equipment includes high resolution color monitors, printers, scanners, video cameras, and digitization tablets. Other equipment includes laptop and notebook computers, such as GRID Pads and Zenith Supersports and Zeos

## VI. LOOK TO THE FUTURE

The future of all three of the efforts described in this report of crop acreage estimation using a regression estimator, vegetative indices, and geographic information systems is bright concerning the technology aspects. The pressure will be on economic factors and showing cost effective improvements or new products in the budget decision time schedules and framework.

As far as the technological aspects, the U.S. Government has recently increased its commitment to future Landsat 6 and 7. The U.S. Government is firmly supportive of the NOAA/AVHRR program. It is currently funded to the year 2005. NASS is firmly supportive of area frame sampling as its statistical foundation to complete universe coverage without duplication in the frame. NASS complements this with list and multiple frame sampling as well. NASS staff are also investigating panel surveys calibrated to the universe as a potential path to reducing total

respondent burden. Geographic Information Systems are proliferating throughout the public and private sectors on a worldwide basis.

The one down side on sensors is that for forecasting and estimating a dynamic event like crop production frequent satellite coverage is required. One Landsat TM or enhanced TM at a time, only gives 16 day coverage. For acreage estimation with the regression estimator, optimum classification windows are often only 30-45 days in length. Usually, that gives only 2 or 3 chances to get data during the optimum window. If those 2 or 3 chances are substantially cloud- covered, then the statistical gains of the regression estimator can drop dramatically. NOAA/AVHRR gives daily coverage but with much different resolution than Landsat TM or SPOT. Thus, it enables large scale looks at the vegetative indices across time but doesn't provide a vehicle for estimating acreage accurately compared to ground-gathered data systems. Perhaps some private sector systems could be developed to better meet agriculture's needs.

The challenge will be to speed up the R&D process as much as possible to evaluate if cost beneficial application of these various technologies is appropriate under most likely declining budgets. Substantial progress has been made but work remains. The U.S. and other government commitment to space borne sensors seems to be at a quite healthy stage. The next 5 - 10 years will be crucial to complete R&D, and to apply the technology where it makes sense in a cost effective manner.

In addition, new sensors such as several nation's radar based systems and NASA's Earth Observing System Data and Information System (EOSDIS) will be new systems of data to evaluate. It is difficult to envision that preciously few research resources in NASS can address new sensors as well as current sensors. NASS staff will observe other efforts such as European and Canadian research on radar systems for agriculture and land cover and NASA research on EOSDIS. Radar sensors overcome the cloud problem but also have different characteristics and require different processing methods. If substantial demonstration of potential cost effective improvements are completed, then NASS research staff would re-evaluate its resource allocation. However, given current resource availability and NASS applications, we will continue to focus on Landsat TM, for crop acreage NOAA/ AVHRR vegetative index for crop condition, and geographic information systems especially related to environmental data such as farm chemical data. In fact, it will be a serious challenge to even address these three applications appropriately under cost and staff constraints.

## VII. ACKNOWLEDGEMENTS

The authors sincerely acknowledge the contributions of many persons associated with this effort. Especially those units and staff listed here: Remote Sensing Section Staff, Technology Research Section Staff, Various Functional Unit Staff Members Throughout the Agency, State Statistical Offices (Arkansas, Mississippi and Louisiana), Agricultural Research Service Staff, USDA Remote Sensing Coordinator, and the Environmental Statistics Staff.

VIII.REFERENCES


Allen, J.D. and Hanuschak, G.A., 1988, "The Remote Sensing Applications Program of the National Agricultural Statistics Service: 1980 - 1987," U.S. Department of Agriculture, NASS Staff Report No. SRB-88-08

Battese, G.E., Harter, R.M. and Fuller, W.A., 1988, "An Error-Components Model for Prediction of County Crop Areas using Survey and Satellite Data," Journal of the American Statistical Association,83(401): 28 - 36.

Bellow, Michael and Graham, Mitchell, 1992. "Improved Crop Area Estimation in the Mississippi Delta Region Using Landsat TM Data," American Society of Photogrammetry and Remote Sensing Convention, Washington, D.C.

Bellow, M.E. and Ozga, M., 1991, "Evaluation of Clustering Techniques for Crop Area Estimation using Remotely Sensed Data," American Statistical Association 1991 Proceedings of the Section on Survey Research Methods, Atlanta, GA, pp. 446 - 471.

Caudill, Charles E. and Hanuschak, G.A., 1983. "Management Issues of Integrating Earth Resourc Satellite Data in to the U.S. Department of Agriculture's Domestic Crop-Area Estimation Program,"Semi-Annual Meeting of the Institute of Management Science/Operations Research Society of America,Orlando, FL.

Cochran, William G., 1977. "Sampling Techniques," New York, NY: John Wiley and Sons, Inc.

Cook, P.W., 1982. "Landsat Registration Methodology Used by U.S. Department of Agriculture's Statistical Reporting Service 1972 - 1982," Washington, D.C.

Craig, Michael E., 1992. "Applications of Advanced Technology for Agricultural Statistics." United Nations, Conference of European Statisticians, Bratislave, Czech and Slovak Federal Republic.

Hanuschak, George, 1977. "Landsat Estimation with Cloud Cover," Proceeding of the 1976 Symposium on Machine Processing Remotely Sensed Data, West Lafayette, Indiana.

Hanuschak, G.A., R.D. Allen and W.H. Wigton, 1982. "Integration of Landsat Data into the Crop Estimation Program of USDA's Statistical Reporting Service 1972 - 1982." Paper presented at the 1982 Machine Processing of Remote Sensed Data Symposium, West Lafayette, IN.

Hanuschak, G.A., and K.M. Morrissey, 1977. "Pilot Study of the Potential Contributions of Landsat Data in the Construction of Area Sampling Frames," U.S. Department of Agriculture, Statistical Reporting Service.

Holko, Martin L., and Richard S. Sigman, 1984. "The Role of Landsat Data in Improving U.S. Crop Statistics," Paper presented at the Eighteenth International Symposium on Remote Sensing of Environment, Paris, France.

Huddleston, Harold F., 1978. "Sampling Techniques for Measuring and Forecasting Crop Yields, "Economics, Statistics and Cooperative Service, U.S. Department of Agriculture, Washington, D.C.

Johnson, R.A. and Wichern, D.W., 1988, "Applied Multivariate Statistical Analysis," Prentice Hall,Englewood Cliffs, N.J.

Ozga, M., Mason, W.W. and Craig, M.E., 1992. "PEDITOR - Current Status and Improvements," in Proceedings of the ASPRS/ACSM Convention, Washington, D.C.

Ozga, M., W. Donovan and C. Gleason, 1977. "An Interactive System For Agricultural Acreage Estimates Using Landsat Data," Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data, West Lafayette, IN.

Sigman, Richard and Gail Walker, 1982. "Use of Landsat for County Estimates of Crop Areas: Evaluation of the Huddleston-Ray and Battese-Fuller Estimators," SRS Staff Report No. AGES 920909, U.S. Department of Agriculture, Statistical Reporting Service.

Stakenborg, Jacques, 1989. "Data Treatment for Crop Statistics," European Communities Joint ResearchCenter. Institute for Remote Sensing Applications Conference on the Application of Remote Sensing to Agricultural Statistics, Varese, Italy.

Winings, Sherman B., 1982. "Landsat Image Availability for Crop Area Estimation." Paper presented at the Eighth International Machine Processing of Remotely Sensed Data symposium for Applications of Remote Sensing, West Lafayette, IN.