SAMPLING METHODS IN AGRICULTURE

R. R. Bosecker, U. S. Department of Agriculture
14th & Independence Avenue, S.W., Washington, D.C. 20250

ABSTRACT The U.S. Department of Agriculture was created in 1862 and the first U.S.D.A. crop report appeared in July 1863. Department statisticians relied on the expert judgment of voluntary crop reports in each county to supply information on crop conditions, production and yields and livestock inventories. These nonprobability measures were converted to official estimates based on relationships to census year or, later, the historic relationships between reports and the final revised estimates based on administrative check data. Probability samples now play the dominant role in the generation of agricultural statistics. Information provided to the public is contained in about 300 national reports and 9,000 state reports issued by NASS each year on more than 150 crop and livestock items, numbers and sizes of farms, farm labor and wage rates, and prices received and paid by farmers. This paper describes past sampling methods and describes current sampling methods used by NASS.

INTRODUCTION Early efforts at determining the nature and extent of agricultural activities in the United States depended upon subjective evaluations by knowledgeable people in various localities. George Washington corresponded with a "purposive sample" of land holders in 1791 for information on farmland prices, commodity prices and crop yields for several areas of the young nation. Today, agricultural statistics are compiled for more than 150 crop and livestock items covering the 50 states through the use of probability samples from area and list sampling frames, the application of remote sensing techniques using satellite data, the collection of data for sample farms via computer-assisted telephone interviewing, and the processing of large quantities of data through a nationwide computer network. The entire national survey process can be accomplished in less time than Washington could send and receive his letters of inquiry. This paper will present the highlights of the changes in sampling methods for agricultural statistics in the United States.

PURPOSIVE (JUDGEMENT) SAMPLING Taking a census was the common procedure utilized by governments for measuring people and product in the early 19th century. In the U.S., the 1840 Census contained agricultural questions for the first time. The decennial census, however, would not provide the data often enough to satisfy the needs of the largely agricultural society. Henry Ellsworth, Commissioner of Patents, used the 1840 census results as a base and, with reports from selected correspondents throughout the States and Territories, issued yearly estimates of agricultural production for 1841 to 1848.

In July 1863, 125 years ago, the newly formed U.S. Department of Agriculture (1862) issued its first crop report. The Agriculture Department again utilized purposive selection procedures with well distributed correspondents. By 1866, a regular reporting series had begun on livestock numbers, monthly crop conditions and final acreage, yield and production estimates for principle crops. "Expert judgement" of local conditions and changes in acreage and production since the previous year was solicited. Annual changes were "benchmarked" against the most recent census results, and estimates were revised upon release of new census counts.

By 1898, there were 10,000 county correspondents and 28,000 township reporters supplying reports to Washington, D.C. In addition, 41 state agents were receiving reports from 7,000 voluntary assistants and providing their assessment on crops, livestock and wage rates to headquarters. In 1905, the Crop Reporting Board was formed as a panel of statisticians for the joint evaluation of the field reports. The monthly condition reports were translated by the Board into crop production forecasts for the first time in 1911.

REPRESENTATIVE SAMPLING The first objective determination of crop acreages began in 1916 with the counting of the number of fields devoted to each crop from a train window during the travels of the statistician in South Carolina. Positive results from this "route sample" of fields led other state statisticians to experiment with counts of fields by crop and to incorporate a measure of field size as indicated by the number of telephone poles opposite each field along predetermined routes. In 1923, the Statistician for Mississippi, D.A. M^{C}Candliss, constructed a "crop meter" for measuring the road frontages of crops by automobile. The total acreage of each crop for the region traveled was then roughly proportional to the ratio of total frontage to total distance covered. By traveling the same route in succeeding years a measure of change in acreage was also readily available.

Objective measurements on growing plants to indicate probable yield began with a plan submitted by Frank Parker, North Carolina office, in 1925 for cotton. Identification of type of crop and

relative size of field along selected routes provided the means for systematic selection of fields along predetermined routes. Preharvest wheat cuttings to estimate yield were begun in 1938. Depending on the density of wheat acreage in an area, intervals of 25 to 200 "crop meter units" of wheat on the right hand side of the road established the sampling interval. (A crop meter unit equaled 1/50 mile.) Metered acreage together with yield per acre indications from plant measurements provided an "objective" estimate of crop production to use in conjunction with the "subjective" evaluations of crop reporters.

Nonprobability samples continued to be the source of data collected both objectively and subjectively. Extensive coverage by local observers throughout the county was relied upon in lieu of the capability to randomly select reporters. In the spring of 1922, while riding to work together, Secretary of Agriculture Henry Wallace and Postmaster General Hubert Work discussed the possibility of using rural mail carriers to distribute questionnaires along their routes for general distribution to farmers. The result was many more thousands of reports pouring into the state statistical offices for tabulation.

The "rural carrier survey" afforded the opportunity to move away from a "judgment" report of change occurring in the respondent's locality to a "factual" report of the data applicable to individual farms. This improvement was especially important to livestock estimates. Farmers and ranchers could now be asked for their inventory of animals as of a specific date for the current and preceding month or year to permit the calculation of percent change based on actual numbers of head on hand. In addition to inventory change ratios, averages per farm could be computed for pigs per litter, milk per cow, chicken and egg production, calf crop, and wool production. Experiments with individual farm reports had begun as early as 1918 and 1919 but the June 1922 Pig Crop report was the beginning of extensive use of individual farm data based on nonprobability rural carrier returns.

PROBABILITY AREA FRAME SAMPLING A major turning point in gaining the capability to select a probability sample of farms in the United States came with the development of an area sampling frame known as the "Master Sample of Agriculture." Iowa State College, now Iowa State University, began development of the Master Sample Frame in 1943 under a cooperative agreement with the Department of Agriculture and the Bureau of the Census. The Bureau utilized a sample from the new frame to collect supplementary information in conjunction with the 1945 Census of Agriculture.

Maps and aerial photographs were used to subdivide the land mass of the United States into area sampling units with identifiable boundaries for enumeration. Three strata were established: (1) incorporated areas, (2) unincorporated areas relatively densely populated, and (3) unincorporated areas sparsely populated (open country). Sample areas (6700 segments) were selected to contain approximately 1/18 of the U.S. land area, 1/18 of the farms (about 300,000) and 1/18 of the rural population. Segments averaged 2.5 square miles but varied greatly in size from one region to another in order to roughly equalize the number of farmsteads in each sample unit. Farmsteads, i.e., farm headquarters, were used to uniquely associate a farm with only one sample unit.

The Master Sample was used extensively in the late 1940's and early 1950's for probability surveys but not as the basis for a recurring statistical program due to lack of funds. In 1954, with the appropriation of funds for methods research and development, a mid-year area frame survey of planted acres and livestock inventory called the June Enumerative Survey was begun on a pilot basis in 10 states (703 sample segments in 100 counties). Research on an end-of-year December Enumerative Survey, primarily for livestock data, began in 1955 using a subsample of area tracts from the June survey. A tract is a parcel of land within a segment belonging to a unique farm operation.

In 1965, the June and December enumerative surveys had been extended to all 48 conterminous states. The area sample consisted of about 17,000 segments and included all or parts of 70,000 farms. Most of the country still relied upon the Master Sample frame for sample segments, but new frames had been constructed for western and northeastern states. These new area frames had utilized stratification based on land-use categories according to intensity of cultivation or urbanization. Visual interpretation of aerial photograph, provided the means to stratify according to land-use. Reductions of 10 to 25 percent in sampling variance were afforded by the new frames. The program to update area frames was made continuous.

Another significant development in the construction of replacement area frames for states occurred in 1977. Hanuschak and Morrisey demonstrated the feasibility of using LANDSAT satellite imagery for area frame stratification by land-use.

Current color satellite images proved to be very beneficial in delineating land-use categories. California, in 1979, was the first state where LANDSAT images were used for restratification and frame construction. Since then, satellite coverage has been obtained for each state receiving an updated area frame.

DOUBLE SAMPLING When the first Earth Resources Technology Satellite (ERTS), later renamed LANDSAT I, was launched in 1972, USDA was ready to research the potential for improving crop acreage estimates. With an enumerated area sample in place to provide "ground truth" information, USDA was in an excellent position to combine the wide area coverage of the satellite images with the detailed crop identification of the sample segments in a double sampling regression estimator.

Research was conducted from 1972 to 1977 to improve selected crop acreage estimates at State, district and county levels. In 1978 the first operational application of satellite data was implemented to enhance end-of-year estimates of corn and soybean acreage estimates for Iowa. The procedure proved useful, especially at State and district levels, and had expanded to eight states and additional selected crops by 1985. In 1988, the operational program has been suspended in favor of devoting the available resources back into research of the new, more powerful satellite sensors.

MULTIPLE FRAME SAMPLING It did not take long during the research phase of area frame sampling to realize that a few large livestock operations could distort survey results by their presence or absence in the sample. Their presence contributed greatly to the sample variance. Therefore, a list of these "extreme operators" was developed independently of the area work. A list sample could then be selected, and the livestock data for extreme operators removed from area samples. This approach was the beginning of multiple frame sampling for agricultural surveys.

An application of the combined use of list and area frame sampling occurred in 1949 for a "Sample Survey of Retail Stores" by the Bureau of the Census. A description of this survey is given in Hansen, Hurwitz and Madow (1953), Sample Survey Methods and Theory. Research supported by USDA at Iowa State University, under the leadership of H.O. Hartley, resulted in theoretical work on multiple frame sampling presented in the 1962 Proceedings of the Social Statistics Section of the American Statistical Association. Multiple frame methods were first evaluated by USDA in Wyoming and Mississippi during 1965. Cooperative pilot studies between USDA and Texas A&M used extended lists of rice and beef producers together with the Texas area frame in 1966-67. Multiple frame studies were also conducted in Illinois, Tennessee, Oklahoma and New Mexico during this same period.

In 1969, four states began a continuous series of multiple frame surveys utilizing extensive lists stratified by size for hog and cattle estimates. By 1974, multiple frame surveys were conducted in 14 states for hog estimates and in 28 states for cattle. Multiple frame surveys that used these extensive list frames provided significantly lower sampling errors than surveys using only small lists of extreme operators.

Various lists of farm operators were compiled and maintained within State Statistical Offices during the 1970's. Multiple frame surveys for rice (1977) and farm grain stocks (1979) were added to those for hogs and cattle. In 1979, work began to develop a comprehensive national list frame system that would standardize the maintenance of names and associated relevant size data for farms and ranches throughout the U.S. By 1981, a national list had been consolidated in one computer system. Maintaining the list frame remained the responsibility of individual State offices. Sampling of the list frame for national multiple frame surveys became centralized in headquarters.

The area frame reporting unit used in conjunction with the Master Sample of Agriculture was based on identification of the "resident farm operator"--those farm operators whose residences were inside the segment boundaries. This reporting unit was the basis for "open segment" estimates using whole farm data. In 1954, with the expanded research program, the "closed segment" approach was introduced which utilized all data physically occurring within the boundaries of the segment. The closed segment estimator generally decreased the wide variation among segment totals that occurred when using whole farm reports for the open segment estimator. Thus, the closed segment estimator usually had smaller sample variances than the open segment estimator. However, the closed segment approach was inappropriate as a means to collect some variables, such as pig crop or calf crop, so both estimators were employed depending upon the item of interest.

The greater share of the total multiple frame sampling errors came from the area frame "nonoverlap" contribution, i.e., farms and ranches not found in the list frame. The number of operators living within sample segment boundaries were

decreasing with the falling number of farms and those who were nonoverlap became a rare occurrence. This led to higher multiple frame variances when the open segment estimator was used for the nonoverlap domain. Therefore, an alternative approach was desired that would: (1) better measure a rare population, (2) reduce the nonoverlap sampling error and therefore the multiple frame sampling error, and (3) utilize data that could be collected on a whole farm basis.

MULTIPLICITY SAMPLING A "weighted segment" estimator had been researched in 1962 during the expansion phase of the area frame surveys. The weighted approach uses entire farm data in conjunction with a weight calculated by dividing the amount of land inside the segment for an operation (tract) by the total number of acres in the farm, i.e., tract acres/farm acres. This is, in fact, a form of multiplicity sampling quite suitable for measuring a rare component of the population.

The multiplicity technique identifies a member of a small subgroup through an association with a larger network. Access through the network results in differential probabilities of selection for reporting units necessitating the weighting of responses by the reciprocal of the multiplicity factor. In agriculture, the land holdings of the farmer or rancher provide a network of acres that may be accessed in whole or in part through area sample segments. (In population surveys, people related or otherwise associated with those in the desired subpopulation provide the network link.) The multiplicity or weighted approach has been used for measuring the nonoverlap domain of multiple frame surveys since the early 1970's.

SUBSAMPLING The nonoverlap domain sample was redetermined each June and December by matching operator names from sample tracts against the list frame from which the list sample had been selected. December tracts were a subsample of those found in the complete enumeration of sample segments in June. Sampling efficiencies were achieved by restratifying June tracts based on size and type before subsampling. Full area frame estimates were generated in June and December for many variables in addition to the nonoverlap contribution to the multiple frame livestock estimates.

September and March multiple frame surveys were specifically for hog estimates and only an estimate for the nonoverlap domain was desired from the area frame. Cost prohibited the subsampling and interviewing of a cross section of all area tracts to redetermine the "current" NOL domain. Until 1981, nonoverlap tracts from June and December were subsampled for the following quarter and revisited with a face-to-face interview to verify and account for the land in the original tract boundaries. The tract was not permitted to change its nonoverlap status regardless of a change in operators. However, the multiplicity weight might change. Compensating errors were assumed for tracts changing operators in the intervening three months.

In 1981, a new sampling procedure for nonoverlap tracts called the "frozen domain" approach was implemented. The multiplicity weight attached to a nonoverlap operator in the base survey (June or December) was "frozen" at the original probability of selection and attached to the name of the operator for subsequent surveys. The area nonoverlap frame for subsampling therefore became another list of names with the associated probability of inclusion from the base survey. This not only required fewer assumptions of compensating error but permitted much cheaper mail and telephone modes of data collection.

INTEGRATED STRATIFIED SAMPLING Stratified sampling has been used for area and list frame sampling since their inception. The area frame was first stratified according to density of dwellings and then type of land-use as discussed earlier. In 1973, in Nebraska, the use of geographic substrata within land-use strata was introduced together with sample selection within substrata by replication (interpentrating sampling). This method continues to the present.

List samples were selected independently by commodity with each operation stratified by a measure of size for the commodity of interest. To minimize panel bias and respondent burden, list samples were selected through replication with partial rotation among replicates between quarters. In the early 1980's separate list samples were drawn for hogs, cattle, sheep, chickens, farm labor, prices, rice, potatoes, farm grain stocks, and cost of production and other economic surveys for the farm sector. Most surveys were on different time schedules throughout the year but when the survey periods coincided the separate questionnaires were pulled together for completion during a single interview.

In 1984 a new design was introduced in three states that integrated the sample selection for all purposes through one stratification of the list frame. Stratum definitions were established based on presence and size of various commodities. Priorities were established

among strata so those farms with rare commodities of interest and those with the largest values for a commodity were in strata having highest priority. Once a sampling unit had been selected for a stratum based on a given commodity, it was ineligible for selection in a lower priority stratum based on the value for another commodity.

In 1986, crop acreage estimates for all crops were included in multiple frame surveys for the first time and integrated sampling was initiated for the 48 conterminous states. Hogs, crops, and grain stocks were put on common, first-of-the-month reference dates in a quarterly survey program. Cattle and sheep were given a separate integrated sample because of different reference dates. The incorporation of all relevant commodities into multiple frame surveys enabled discontinuation of the area frame based December Enumerative Survey after 1986. The June Enumerative Survey provided the frame of nonoverlap tracts for subsampling in subsequent surveys for the following year.

CONCLUSION Today, sampling by the National Agricultural Statistics Service is characterized by the following procedures. An area frame covering the continental United States, stratified by type of land use, is sampled using replicated simple random selection within geographic substrata. The area frame survey in June, continuous in 48 states since 1965, provides closed segment estimates for planted acreage of major crops and for hogs (in conjunction with an extreme operator list sample) with acceptable precision levels, e.g., 2 percent coefficients of variation (C.V.) nationally and 5 to 10 percent for major states. In addition, the area survey defines the nonoverlap domain in June and provides the means to subsample nonoverlap tracts for subsequent multiple frame surveys.

List frames are generated from a national list of known farm operators with associated classification data for type and size of commodities produced. Multiple frame surveys now collect data for several commodities simultaneously. A stratified random sample is selected from a list of farms having the variables of interest. Stratification is based on measures of size and frequency of occurrence for selected items. Highest priority is given to farms having large values for those commodities least widely produced.

Multiple frame surveys are conducted quarterly for hog, crops and farm grain stocks data, semi-annually for cattle, sheep and goats, and annually for cost of production and farm costs and returns information. A separate multiple frame survey supplies farm labor data quarterly.

In addition to the multiple frame area and list probability surveys, a large scale mailed nonprobability survey continues to be conducted in October-November. This survey relies upon change indicators and means per farm for which a long history has been developed. It is used to assist in allocating State level estimates to the counties within the state.

Not all surveys are from producers. Price data, important since Washington's letters of inquiry, are generally obtained through samples of buyers of agricultural products and sellers of goods and services to farmers and ranchers. Extensive lists, complete for some commodities, are compiled for sampling buyers to determine average prices farmers receive for their products.

To determine prices paid by farmers, those farm operators selected in the multiple frame survey for farm costs and returns are asked to provide the names of firms patronized and the types of items purchased. Sampling frames are constructed from this information and sample firms provide quantity sold and price information for specified items.

Objective and subjective yield forecast indications continue to be available. The objective yield plant counts and measurements had been given a probability sample basis for selection with the creation of the area frame. Some objective yield programs utilize the multiple frame sample for field selection. Subjective yield reports continue to rely upon a panel of crop reporters providing monthly conditions reports and expected yields. These voluntary reporters are part of a continuous chain that extends throughout the 125 year history of USDA statistics.

REFERENCE

1. American Farm Economics Association. Journal of Farm Economics, Vol. 21: November 1939. This issue is devoted to articles commemorating a century of agricultural statistics.

2. Bosecker, R.R. "List Dominant and Frozen Domain Procedures," Staff Report, National Agricultural Statistics Service, July 1984.

3. Hanuschak, George and Katheleen Morrissey. "Pilot Study of the Potential Contributions of Landsat Data in the Construction of Area Sampling Frames," Staff Report,

National Agricultural Statistics Service, October 1977.

4. Hendricks, Walter A. Theoretical Aspects of the Use of the Crop Meter, Agricultural Marketing Service U.S. Department of Agriculture and Works Projects Administration of the City of New York, No. 2, 1942.

5. _____. "The Relative Efficiencies of Group of Farms as Sampling Units", Reprint, Journal of the American Statistical Association, Vol. 39: pp 366-376, 1944.

6. Hendricks, W.A., D. T. Searles and D. G. Horwitz. "A Comparison of Three Rules for Associating Farms and Farmland With Sample Area Segments In Agricultural Survey," Research Triangle Institute, contractual study for Bureau of the Census, 1954 and 1959 Census of Agriculture sample segment data.

7. Houseman, Earl E., and T. J. Reed. "Application of Probability Area Sampling to Farm Surveys," Agricultural Marketing Service, U.S. Department of Agriculture, 1954.

8. Houseman, Earl E., and Joseph A. Becker. "A Centenary Profile of Methods for Agricultural Surveys,: The American Statistician, pp. 15-21, April, 1967.

9. Jessen, Raymond J. Statistical Investigation of a Sample Survey for Obtaining Farm Facts, Iowa Agricultural Experiment Station, Research Bulletin 304, 1942.

10. King, Arnold J., and Dale McCarty. "Application of Sampling to Agricultural Statistics with Emphasis on Stratified Samples," Reprint, Journal of Marketing, pp. 462-474, 1941.

11. King, Arnold J., Dale E. McCarty, and Miles McPeek. An Objective Method of Sampling Wheat Fields to Estimate Production and Quality of Wheat, U.S. Department of Agriculture in cooperation with Kansas, North Dakota, and Iowa Agricultural Experiment Stations. Technical Bulletin No. 814, 1942.

12. King, Arnold J., and R. J. Jessen. "The Master Sample of Agriculture," Reprint, Journal of the American Statistical Association, vol. 40: pp. 38-56, 1945.

13. Nealon, J. P. "Review of the Multiple and Area Frame Estimators," Staff Report, National Agricultural Statistics Service, March 1984.

14. Newell, S.R. Federal-State Crop and Livestock Reporting Service: A chronology of Development and Progress, 1866-1966, Statistical Reporting Service, U.S. Department of Agriculture, 1966.

15. Sarle, Charles F. The Theory of Sampling as Applied to Crop Estimating, Bureau of Agricultural Economics, U.S. Department of Agriculture, 1929. Issued for the use of the Staff of the Division of Crop and Livestock Estimates.

16. _____. The Possibilities and Limitations of Objective Sampling in Strengthening Agricultural Statistics, Research Conference, Cowles Commission for Research in Economics, Colorado Springs, Colorado, July 3-28, 1939.

17. U.S. Department of Agriculture. Statistical Reporting Service. The Story of the U.S. Agricultural Estimates. Miscellaneous Publication No. 1088, 1969.

18. _____. Scope and Methods of the Statistical Reporting Service. Miscellaneous Publication No. 1308, September 1983.