Utilizing an Alternative Sampling Frame to Produce Agricultural Survey Indications

Wendy J. Barboza, Mark Harris

USDA, National Agricultural Statistics Service, Research and Development Division

3251 Old Lee Highway, Room 305 Fairfax, VA 22030

wendy_barboza@nass.usda.gov, mark_harris@nass.usda.gov

Abstract: The U.S. Department of Agriculture's (USDA) National Agricultural Statistics Service (NASS) conducts hundreds of surveys every year and prepares reports covering virtually every aspect of U.S. agriculture. NASS maintains a list frame containing names, addresses, telephone numbers, and other descriptive data on producers (and agribusinesses) and an area frame covering all land area in the U.S. To form multiple frame survey indications, data collected from the list frame sample are combined with data collected from the area frame operators who are not on the list frame. In this respect, the area frame accounts for the incompleteness of the list frame. This methodology ensures that every producer has a chance of selection. For many years, NASS has partnered with USDA's Farm Service Agency (FSA) to use their data as an administrative data source since most producers report their planted crop acreages to FSA on an annual basis. Starting in December 2006, NASS initiated an operational pilot program to employ using FSA administrative data as a sampling frame in the state of Nebraska. The goals of this new methodology were to reduce respondent burden, reduce data collection costs, and improve survey indications. After two years, the operational pilot program was discontinued because the objectives were not sufficiently achieved. However, various operational programs within NASS continue to use FSA administrative data, just not as a sampling frame. This paper provides an overview of NASS utilizing an alternative sampling frame to produce agricultural survey indications.

Key words and phrases: sampling frame, agricultural survey indications, NASS, FSA, QAS, NOL

Introduction

The U.S. Department of Agriculture's (USDA) National Agricultural Statistics Service (NASS) conducts hundreds of surveys every year and prepares reports covering virtually every aspect of U.S. agriculture. NASS maintains a list frame containing names, addresses, telephone numbers, and other descriptive data on producers (and agribusinesses) and an area frame covering all land area in the U.S. To form multiple frame survey indications, data collected from the list frame sample are combined with data collected from the area frame operators who are not on the list frame. In this respect, the area frame accounts for the incompleteness of the list frame. This methodology ensures that every producer has a chance of selection.

For many years, NASS has partnered with USDA's Farm Service Agency (FSA) to use their data as an administrative data source since most producers report their planted crop acreages to FSA on an annual basis. Prior to 1997, NASS researched using FSA data as a sampling frame and administrative data source. After passage of the Federal Agriculture Improvement and Reform Act of 1996 (the 1996 Farm Bill), which reduced the emphasis on producers reporting to FSA, NASS temporarily stopped working on this issue. When the Farm Security and Rural Investment Act of 2002 (the 2002 Farm Bill) required producers participating in the FSA program to report all crops, NASS once again pursued examining different alternatives of using FSA administrative data.

The Quarterly Agricultural Survey (QAS) is conducted once every three months in March, June, September, and December. Data collected by the QAS are used to set national- and state-level commodity estimates for planted acres, harvested acres, production, and on-farm grain stocks. In December 2004, the QAS was utilized to conduct a research study in two states, Nebraska and Minnesota, to evaluate the possibility of using FSA administrative data as a sampling frame (instead of the NASS list sampling frame). Two additional states, North Carolina and Oregon, were added to the research study in June 2005. The survey indications and coefficients of variation produced from the research studies were compared to those calculated from the operational program. The evaluation showed mixed results; the performance of the survey indications and coefficients of variation varied by state and commodity.

NASS subsequently made a decision to initiate an operational pilot program in one state. There was sufficient motivation for using FSA administrative data as a sampling frame in terms of respondent burden reduction, operational efficiency, and quality of survey indications. NASS' target population is small compared to those of many other government agencies, resulting in respondent burden being more of a concern with NASS surveys than elsewhere. With the limited population size, NASS often surveys the same operators multiple times during its annual survey cycle. Consequently, the agency wanted to reduce respondent burden for producers who had already reported their planted crop acreages to FSA or who were being contacted multiple times by various organizations. At the same time, NASS hoped to reduce data collection costs and improve survey indications.

In November 2005, an intra-agency team was established to identify all of the surveys impacted by this decision and to prepare the specifications for integrating the methodology into these surveys. Beginning in December 2006, NASS started using FSA administrative data as a sampling frame in the state of Nebraska. After two years, the operational pilot program was discontinued because the objectives were not sufficiently achieved. However, various operational programs within NASS continue to use FSA administrative data, just not as a sampling frame.

This paper provides an overview of NASS utilizing an alternative sampling frame to produce agricultural survey indications. Although there were a number of surveys affected by the new methodology, this paper will only focus on the QAS. The methodological differences will be covered as well as updates made to the original survey methodology in addition to the reasons for discontinuing the operational pilot program.

Methodology for QAS Using the NASS Sampling Frame

Using the operational NASS sampling frame, the survey cycle for the QAS starts in June (i.e., the base month) and ends in March of the following year. Multiple frame survey indications are produced by combining data from the list frame and area frame samples.

As stated earlier, NASS maintains a list frame containing names, addresses, telephone numbers, and other descriptive data on producers. For the list frame, the sampling unit is a producer on NASS' list and the reporting unit is any farm associated with the producer. So, when a sampled producer is involved in multiple operating arrangements, separate questionnaires are completed for each one. In early spring, the list frame population is "frozen" to select the sample. Strata are defined based on the control data (i.e., previously reported data) for the records. These strata are different for each state, and are not design strata. Instead, they are only used to define the population, identify prob-1 (i.e., certainty) records, and adjust for nonresponse. In 2005, strata in Nebraska were formed using total cropland, calculated cropland, on-farm grain storage capacity, and hay acres.

Reduced-list sampling methodology is used to define the list population; this means that records with small amounts of cropland or capacity are excluded from the sampled population and are represented by the area frame component. Records in the "large" strata are identified as prob-1 records and are included in the sample every quarter. These records are not eligible for nonresponse adjustment and must be manually imputed for refusals and inaccessibles. In order to target a multitude of commodities, Multivariate Probability Proportionate to Size (MPPS) sampling (Kott and Bailey, 2000) is used to sample the remaining reduced-list population and there are three components: general, row crops, and small grains. During sample selection, the control data are used to target the commodities of interest for each of the three components. As with the strata formation, the targeted commodities are different for each state. In 2005, the targeted commodities in Nebraska were calculated cropland, on-farm grain storage capacity, and reported cropland for the general sample; alfalfa hay, corn, dry beans, garbanzo beans, all hay, calculated row crop acres, other hay, sunflowers, soybeans, proso millet, and sorghum for the row crops sample; barley, oats, rye, calculated small grain acres, and winter wheat for the small grains sample. Once the samples for the three components are identified, the records are assigned to one of three replicates. The component and replicate number determine which quarters the sampled record is interviewed. This rotation scheme is shown in the below table.

QAS Rotation Scheme				
Sample Component	Replicate Number			
	June	September	December	March
General	1	1, 2, 3	1, 2, 3	2, 3
Small Grains	1, 2	1, 2, 3	1, 2, 3	2, 3
Row Crops	1	1	1, 2, 3	2, 3

After the list sample is selected, the records within each quarter are calibrated to state-level commodity totals based on the control data of the reduced-list population. This calibration is performed to adjust for procedures related to the implementation of MPPS sampling. After this procedure, each record is assigned an adjusted sampling weight. Note that a record is in sample for multiple quarters, but the sampling weight is different for each quarter.

Records sampled from the list frame are combined with area frame records which are not represented by the list frame component; these area frame records are referred to as the Not-On-List (NOL) component. In June, NASS conducts an agricultural survey of geographic segments (i.e., areas of land approximately 1 square mile in size) sampled from the area frame. For the area frame, the sampling unit is a segment and the reporting unit is an area of land inside the segment that is operated under one type of arrangement. The reporting unit, referred to as a tract, may consist of agricultural land, non-agricultural land with potential, or non-agricultural land. Farm-level data are collected for the agricultural land tracts. These tracts are matched against the reduced-list frame population based on name/address information and the non-matches comprise the NOL component. In June, the sampling weight for NOL agricultural tracts is adjusted by a tract-to-farm weight and the farm-level data are combined with the list frame records to produce multiple frame survey indications. For other quarters, the NOL agricultural land tracts and tracts classified as agricultural land with potential are post-stratified based on data collected in June. A sample is selected and the original sampling weight is adjusted accordingly. These NOL sampled records are interviewed every remaining quarter (i.e., September, December, and March) using the same questionnaire as the list frame records to collect farm-level data.

Original Methodology for QAS Using the FSA Sampling Frame

Using the FSA sampling frame, the survey cycle for the QAS started in December (i.e., the base month) and ended in September of the following year. The reason for changing the base month from June to December was because most producers do not report their planted crop acreages to FSA until later in the calendar year. Similar to using the NASS sampling frame, multiple frame survey indications were produced by combining data from the list frame and area frame samples.

The FSA list population was constructed using three files: the FSA Geographic Information System (GIS) data layer, the FSA 578 administrative data file, and the FSA name/address file. The FSA GIS data layer displayed the geographic boundaries for all FSA farms; this file was also used to calculate the total acres of the FSA farm. The FSA 578 administrative data file contained the current-year planted crop acreages by FSA farm number for producers who reported to FSA. The FSA name/address file contained name and address information of owners, operators, and others associated with the FSA farm. In August, NASS received the FSA files and the list population was considered "frozen" to select the sample. The first two files were matched together using state, county of administration, and FSA farm number. The FSA GIS data layer was considered the primary file; in other words, records in the FSA 578 administrative data file which did not match to the FSA GIS data layer file were discarded from the list population. Records on the FSA GIS data layer file which did not match to the FSA 578 administrative data file contained data for total acres only. For the list frame, the sampling unit was an FSA farm number and the reporting unit was the parcel of land identified by the geographic boundaries. So, if the sampled FSA farm number changed, the questionnaire was completed based on the geographic boundary associated with the original FSA farm number. Similar to using the NASS sampling frame, strata were defined based on the administrative data for the records and were used to define the population, identify prob-1 records, and adjust for nonresponse. In 2005, strata in Nebraska were formed using total land (instead of total cropland), calculated cropland, and hay acres. On-farm grain storage capacity was not used because it is not reported to FSA. Usually, this target variable is used to obtain data for crops stored on the operation (i.e., stocks).

Reduced-list sampling methodology was not used to define the FSA list population (i.e., all FSA farms were included). Records in the "largest" stratum were identified as prob-1 records and were included in the sample every quarter. These records were not eligible for nonresponse adjustment and were manually imputed for refusals and inaccessibles. The remaining list population was sampled using MPPS sampling and there were three components: general, row crops, and small grains. During sample selection, the control data were used to target the commodities of interest for each of the three components. The targeted commodities in Nebraska were the same as those used for the NASS sampling frame with the exception of using total land instead of total cropland and the unavailability of on-farm grain storage capacity. Once the samples for the three components were identified, the records were assigned to one of three replicates. The component and replicate number determined which quarters the sampled record was to be interviewed. The rotation scheme was unchanged from the traditional approach shown earlier. After the list sample was selected, the records within each quarter were calibrated to state-level commodity totals based on the control data of the list population and assigned an adjusted sampling weight. The records in the list sample were then matched to the FSA name/address file. Since multiple names could be associated with the FSA farm, the primary contact was determined by ranking the names based on whether the person was listed as an owner, operator, other, or combination thereof.

Records sampled from the list frame were combined with the NOL component. The NOL component was determined by overlaying the FSA GIS data layer with the NASS GIS data layer for the geographic segments sampled from the area frame in June. Any areas within the segments that were not covered by the FSA GIS data layer were identified, "digitized", and compared to the tract-level data collected in June. The NOL component consisted of all digitized areas within a segment which were classified as an agricultural land tract or a non-agricultural land with potential tract. For the NOL component, the sampling unit and the reporting unit were the same and were referred to as a parcel of land whether it was an entire tract or a "partial" tract (i.e., areas of land within the tract). These NOL parcels of land were assigned the sampling weight from the June Agricultural Survey (since all were selected for sample) and interviewed every quarter using the same questionnaire as the list frame records to collect parcel-level data.

Using the FSA sampling frame, as mentioned above, the reporting unit for both the list and area frame components was a parcel of land. A map was included on the front of the questionnaire as a visual aid to assist the respondent in recognizing the FSA farm number that was selected from the list frame component or the NOL parcel of land identified from the area frame component.

Updates to Methodology Using the FSA Sampling Frame

A detailed analysis was performed on the results from the December 2006 QAS. The results using the FSA sampling frame suggested that survey indications for minor commodities such as sorghum, dry beans, and sunflowers were more precise; results for major commodities such as cropland, corn, and soybeans were mixed; and results for stocks were extremely unfavorable. Another detailed analysis was performed after the March 2007 QAS was conducted. The results suggested that survey indications for soybeans and sorghum were more precise; corn, winter wheat, and all hay were less precise; and dry beans and sunflowers were mixed. Again, the survey indications for stocks were extremely unfavorable. A major issue mentioned in both analyses was that, unlike the NASS sampling frame, a valid zero could not be differentiated from missing data. When FSA 578 administrative data were not present for a commodity, the producer could have not grown the commodity (i.e., a valid zero) or grown the commodity but not reported it to FSA. There was not sufficient time to analyze the results from June 2007 and September 2007 because any methodological

changes had to be specified and implemented before the sample needed to be selected, but the results from December 2006 and March 2007 provided sufficient information to pursue several improvements.

First, although the FSA 578 administrative data were more recent than NASS control data, the geographic area being covered was significantly less since an FSA farm was smaller in size than a NASS-defined farm. In Nebraska, on average, there were 2.4 FSA farms for every NASS-defined farm. The quarterly sample size was increased to increase the amount of land covered. This change was accomplished by modifying the rotation scheme, which is shown below. In order to have better control over the quarterly sample size, the number of replicates was changed from 3 to 7 and the same rotation approach was used for each component. In an attempt to minimize respondent burden, the total sample size for all four months was kept the same as the original approach. Although the same number of producers would be contacted, they would be contacted more often than before.

Modified QAS Rotation Scheme				
Sample Component	Replicate Number			
	December	March	June	September
General, Small Grains, Row Crops	1,2,3,4,5,6,7	1,2,3,6	1, 3,5,6	2,3,5,6,7

Second, on-farm grain storage capacity was added into the sampling process. This change could be implemented because NASS internally developed a file of FSA farm numbers with control data for on-farm grain storage capacity. Unfortunately, only a partial file could be created because of time constraints. This partial file was put together using several sources, first of which was previously reported survey data. Any FSA farm numbers that had reported data for on-farm grain storage capacity during the QAS in December, March, or June were identified. The second source was the NASS list frame. A file of name and address information for records on the NASS list frame with on-farm grain storage capacity was created. Due to timing, only records on the NASS list frame with 5,000 or more bushels of on-farm grain storage capacity were included in this process. This file was then matched to the FSA data files. When only one FSA farm matched to one NASS record, the FSA farm number was assigned the control data value from the NASS list frame. The FSA farm numbers for the remaining matches (i.e., multiple FSA farms matched to one NASS record) were identified on the FSA GIS data layer and the control data value from the NASS list frame was transferred over. This revised FSA GIS file and then overlaid with satellite data. The FSA farm's on-farm grain storage capacity was estimated by visually inspecting the circular bins of all FSA farms associated with the NASS record and assigning values which would sum to the control data value. Again, due to time constraints, the manual process focused on determining the control data values for FSA farms that had the largest NASS grain storage values associated with them. At the end of the process, the resulting partial file of FSA farms numbers with grain storage was matched to the new FSA list frame population and on-farm grain storage capacity was added to the other control data for matches.

Third, some producers had difficulty reporting for the parcel of land even though a map was included on the front of the questionnaire as a visual aid. This problem was mentioned by enumerators during the data collection phase. In an effort to address this issue and reduce respondent confusion, boundaries and information for the legal description and minor civil divisions (MCDs) were added to the map. Previously, the map only showed the road names.

Lastly, the logic for utilizing the FSA 578 administrative data for various survey procedures was updated. As stated earlier, the FSA 578 administrative data file contained the current-year planted crop acreages by FSA farm number for producers who reported to FSA. In the original methodology, when administrative data were not present for a commodity, the planted crop acreage was considered missing when it could have been a valid zero (i.e., the producer did not grow the commodity). The missing planted crop acreages were replaced with a zero when it could be definitely determined that the producer did not have a particular commodity. This new logic would potentially improve various survey procedures as well as certain ratios being generated as part of the survey indications.

Reasons for Discontinuing the Operational Pilot Program

After the above methodological changes were implemented, another detailed analysis was performed on the results from several QAS quarters. However, this analysis was extended to include all of the original goals rather than only the precision of the survey indications. As stated earlier, the goals of the new methodology were to reduce respondent burden, reduce data collection costs, and improve survey indications.

Respondent burden: The average time to complete an interview using each sampling frame was compared for corresponding quarters. There was no evidence the average interview time for the FSA sampling frame was less than the NASS sampling frame. When considering other issues, respondent burden actually increased using the FSA sampling frame. For example, a producer was asked to complete multiple questionnaires more often because it is more common for a producer to be associated with multiple FSA farms than multiple NASS-defined farms. Also, some producers had difficulty reporting for a specific FSA farm but nearly all producers recognize the land associated with their NASS-defined farm.

Data collection costs: The average cost per sample using each sampling frame was compared for corresponding quarters. This cost represented the total number of hours charged by enumerators during the data collection phase. There was no evidence the average cost per sample for the FSA sampling frame was less than that of the NASS sampling frame. Although data collection costs were not different, the overall cost to the agency was actually higher using the

FSA sampling frame. This occurred because the FSA sampling frame required NASS staff to perform extensive additional work related to operational procedures (e.g., processing and manipulation of files, survey preparation and coordination, etc.). Thus, the increase in the total number of staff hours resulted in a higher overall cost than using the NASS sampling frame. Over time, this higher overhead cost may have been reduced as the process was refined and streamlined.

Survey indications: The survey indications for acreage, yield, and production were generally more precise using the NASS sampling frame. It was speculated that these results were due to the FSA sample covering less total land than the NASS sample, even though the FSA control data were more current. This hypothesis is probably true since the overall sample size was kept constant due to respondent burden and costs, but the FSA sampling frame was approximately 2.4 times larger than the NASS sampling frame. In addition, the precision levels of survey indications for stocks were extremely unfavorable when using the FSA sampling frame, even after the methodological changes were implemented. It was speculated that the survey indications for stocks would still be unsatisfactory even if control data for on-farm grain storage capacity was available for all FSA farms. The rationale for this hypothesis is difficult to explain and beyond the scope of this paper.

In conclusion, the analysis revealed that the objectives were not sufficiently achieved. After two years, the operational pilot program of utilizing an alternative sampling frame in Nebraska was discontinued. However, various operational programs within NASS continue to use FSA administrative data, just not as a sampling frame.

References

Kott, P.S. and Bailey, J.T., (2000), "The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling," Proceedings of the International Conference on Establishment Surveys II, 269-278.