AN APPLICATION OF MULTIPLE LIST FRAME SAMPLING FOR MULTI-PURPOSE SURVEYS

Jeffrey T. Bailey USDA-NASS, Phillip S. Kott USDA-NASS
Jeffrey T. Bailey, National Agricultural Statistics Service, 4818 South Bldg., Washington, D.C. 20250

KEY WORDS: Multiple List Frames, Poisson Sampling, Calibration.

I. Introduction

Many surveys are multi-purpose and are designed to gather information on various related items. Determining a sample design to efficiently and effectively address estimation needs for multiple items is often a challenging task. This process can, however, be expedited with the utilization of faster computers and new estimation methods. This paper examines an estimation strategy that selects a Poisson sample from multiple list frames and uses calibration estimation.

The estimation strategy is applied to the National Agricultural Statistics Service's (NASS) Crops Survey (CS). The CS produces estimates for acreage, yield, production and stocks for multiple crops. The survey instruments solicit information during the quarter that the information is pertinent. For example, questions about acreage, yield and production are asked seasonally in conjunction with a given crop's planting and harvesting dates, while questions about stocks for major crops are asked each quarter. In addition, any design for the CS must also address the inclusion of uncommon, yet important commodities.

Beginning in June 1997, the Minnesota CS is being conducted using samples selected from both the old and new designs. In future discussions Minnesota's particular design will be used when examining the new design or comparing it to the old design.

II. Current Sampling Design

The current CS utilizes a common priority stratified design which is unique for each state. Under this design, population units that are largest in size in general categories, such as cropland or capacity, are grouped to insure inclusion or a high probability of selection. This grouping process is continued until the smallest units are grouped and probabilities of selection are assigned accordingly. This design works well when there are only a few commodities and the resulting strata are fairly homogenous. A commodity that is uncommon but important, is often grouped and placed in a strata priority order that would insure representation and insure a minimum number of samples from this population. Unfortunately, this design makes it difficult to control sample sizes for individual crops and to target the samples for particular crops in different quarters. The samples are replicated so that a subset of replicates can be used in each quarter. The number of replicates used each quarter varies depending upon survey needs.

Three different non-overlapping samples are needed in the crops estimating program:

1. The *CS is the principal sample* to estimate crop acreage, yield, production, and stocks. This sample is used each quarter in March, June, September, and December. Survey questions change quarterly depending on the crop information that is pertinent for that time period. Sample replicates are rotated to have controlled overlap from previous quarters and to reduce respondent burden.

2. A *row-crop yield sample* to estimate yield information for row crops during the months of August through November. The sample is screened in June as part of the CS to identify operations with the crops of interest for this yield survey.

3. A *small-grains yield sample* to estimate yield information for small grain crops during the months of May through September. The sample is screened in March as part of the CS to identify the operations with the crops of interest for the small-grain yield survey.

There are several problems with the current design which can be summarized by the need to target samples for specific crops. The new design should improve the estimation of rare crops, while targeting specific crops only in the time periods the information is needed.

III. The New Sampling Design

The new design will draw three non-overlapping samples, the principal CS sample and two yield samples. In the new design, the principal CS sample is comprised of a number of "components" drawn from *partially overlapping* frames. Each frames target a specific population: general, row crops, small-grain crops, or specific individual crops. Inclusion of a producer in any frame depends on the presence of specific control data in the NASS list frame database associated with the operation. Table 1 shows each frame and the quarter that a sample component will be used.

Table 1: Population and Period of Interest

Quarter	General	Row Crop	Small Grain	Crop Specific (Potato)
March	x	x	x	x
June	x	x	x	
September	x		x	x
December	x	x	x	

A *general frame* contains all crop farms known to NASS in a state. The control value used is a capacity equivalent which is the maximum of storage capacity or total cropland acres times 60. A general sample component to represent all crop farms is drawn from this frame for use in every quarter.

A *row-crop frame* contains farms that have been identified as row crop producers for such crops as soybeans and corn. The row-crop sample component selected from this frame is surveyed in March, June and December. The *row-crop yield sample* is also drawn from this frame.

A *small-grains frame* contains farms that have been identified as small grain producers for crops such as wheat and barley. The small-grains sample component is selected from this frame and is used in March, June, September and December. The *small-grains yield sample* is also drawn from this frame.

Crop specific frame(s) identifies a specialty crop of interest that only needs to be surveyed in particular quarters. For example, in Minnesota, potatoes is a specialty commodity for which estimates are needed in the quarters of June and December.

Determining the Selection Probabilities for Sample Components

For each sample component, a minimum sample size is specified. In addition, minimum sample sizes are also required for farms believed to have particular rare crops. For example, the row-crop yield sample in the test state of Minnesota has a minimum size of 500. The sample is also required to produce 100 sunflower farms and 100 edible bean farms.

In other words, each sample component has a number of target variables. For the row-crops yield sample, these target variable are sunflowers, edible beans, and total row crops. To determine the farm (unit) selection probabilities for a sample component that would likely meet all targets and be fairly efficient, the following procedures are employed:

A. Ascertain the minimum sample size, n_j, required for each target variable j.

B. Let x_{ij} denote the measure of size (control value) for the target variable j in farm i.

C. $$Let\ \pi_i = \max\left(n_1 * \frac{\sqrt[3/4]{x_{i1}}}{\sum_{i=1}^{N} \sqrt[3/4]{x_{i1}}}, \ldots, n_J * \frac{\sqrt[3/4]{x_{iJ}}}{\sum_{i=1}^{N} \sqrt[3/4]{x_{iJ}}} \right)$$

D. Let $\pi_{iF} = \min\{L, \pi_i\}$ be the selection probability attached to farm i in the frame from which sample component F is to be drawn; L is the predetermined limit on π_{iF}.

E. Assign each farm a permanent random number(PRN) from the interval [0, 1) and select a farm if its random number is less than π_{iF}.

Sequence in Selecting the Sample:

As described earlier, three non-overlapping samples (the CS and two yield samples) are desired. Sequential sampling with adjustments to the probabilities for the selection of the second and following samples accomplish this goal. The sequence is described below:

1. A Poisson sample for row-crops yield is selected from the row-crops population.

2. A Poisson sample for small-grain crops yield is selected from the small-grains population excluding from the population any units that had been selected for row-crops yield.

3. CS sample components from each of the four populations is selected *independently* of the yield samples; the same random number is assigned to a farm for each of these four sample components.

4. Farm records chosen for either yield sample are removed from the combined CS sample in Step 3; the remainder are randomly assigned into one of three (3) replicates.

Sample selection for the 1997 Minnesota new design was conducted in the four step sequence outlined above with the following sample sizes and probability limits described below:

Step 1 - Selecting the sample for row-crops yield

i. A total sample of about 500 was desired. Commodities with specified minimum samples were sunflowers(n=100) and dry edible beans(n=100).
ii. Probability limit L=1/3.
iii. Let farm i's probability of selection for the row-crops yield sample (computed using A through E) be denoted as π_{ir}

Step 2 - Selecting the sample for small-grain yield

i. A total sample of about 500 was desired. Commodities with specified minimum samples were winter wheat(n=120), durum wheat(n=50) and barley(n=100).
ii. Probability limit L=1/3.
iii. The overall probability of selecting farm i for the small-grains yield survey was determined using A through E and will be denoted as $\pi_{is'}$. To achieve this probability, the sample was selected among farms not sampled for the row-crops yield sample with these conditional (on not being selected for the small-grains yield sample) probabilities: $\pi_{is} = \pi_{is'}/(1 - \pi_{ir})$.

(Note: The probability of farm i being in either yield survey is $\pi_{iY} = \pi_{ir} + \pi_{is}$ which cannot exceed 2/3.)

Step 3 - Selecting the CS Sample

Sample components were drawn from four different populations with no probability limit.
i. General (n=800)
ii. Row crops (n=1100) with sunflowers (n=100), dry beans (n=150) and flax(n=80)
iii. Small grains(n=680) with rye(n=150), winter wheat(n=150) and durum (n=60)
iv. Potatoes (n=100 for June; n=50 for December)

GSRP denotes the set of farms selected from at least one of the four frames when n =100 for the sample component from the potato frame, and GSRP* denotes the set of farms selected from at least one of the four frames when n =50 for the sample component from the potato frame. GSR denotes the set of farms selected from either the general, small grains, or row-crops frame (or any combination of the three). GR (general and row crops) and GS (general and small grains) are defined similarly.

Let farm i's probability of selection from frame F be denoted by π_{iF}. $\pi_{iGSRP} = \max\{\pi_{iG}, \pi_{iS}, \pi_{iR}, \pi_{iP}\}$ is utilized to denote farm i's probability of selection into GSRP. Analogously , π_{iGSRP*}, π_{iGSR}, π_{iGR} and π_{iGS} are defined.

Step 4 - Creating Replicates

Those farms from GSRP that are already in one of the yield samples are removed and the remaining farms are then randomly assigned to one of three (3) replicates. Replicates for GSRP*, GSR, GR and GS farms are assigned in a like manner.

Probabilities of Selection for the Surveys combining the Samples

For June: All farms in Replicate 1 of GSRP and all GSRP farms in the potato frame are enumerated along with all farms in the row-crops yield survey. The selection probability for a chosen non-potato-frame farm i is $(1 - \pi_{iY})\pi_{iGSRP}/3 + \pi_{ir}$. For a potato-frame farm, it is $(1 - \pi_{iY})\pi_{iGSRP} + \pi_{ir}$.

For September: All farms in GS from any of the three replicates are enumerated. The selection probability for a chosen farm i is $(1 - \pi_{iY})\pi_{iGS}$.

For December: All GSRP* farms from any replicate are enumerated. The selection probability for a chosen farm is $(1 - \pi_{iY})\pi_{iGSRP*}$.

For March 1998: All GSR farms in Replicate 2 or 3 and all farms in the small-grain yield survey will be enumerated. The selection probability for a chosen farm i is $(1 - \pi_{iY})\pi_{iGSR}(2/3) + \pi_{is}$.

Note on the use of Poisson PRN sampling

That sample sizes determined by Poisson sampling are random is merely a disadvantage that is compensated by the use of a calibration estimator. In addition, Poisson PRN sampling expedites the determination of the probability of selection for whatever set of sample components are desired in any particular quarter. For example, the September Survey requires samples from only the general and small-grain samples. The probability of a farm's being in both the general and the small-grains sample is determined by:

π_{iGS} = P(i selected for small-grains sample *or* i selected for general sample),
= π_{iG} + π_{iS} - P(i selected for small-grains sample *and* i selected for general sample).

This is not always easy to calculate in general. In our Poisson PRN design, however, with the same random number for farm i applied to every frame, π_{iGS} is simply $\max\{\pi_{iG}, \pi_{iS}\}$. The new design minimizes π_{iGS}, because P(i selected for small-grains sample *and* i selected for general sample) is maximized at $\min\{\pi_{iG}, \pi_{iS}\}$.

Amrhein, Hicks, and Kott (1996) suggests that π_{iGS} be set to the $\max\{\pi_{iG}, \pi_{iS}\}$ and that a systematic probability proportional to "size" sample then be selected. The problem with this approach is its inability to determine for which sample components a farm is selected. This failing could lead to an unnecessary burden on potato-only farms in September and March.

IV. Calibration

Calibration is the final step in determining the weight of each farm. Calibration of the weights provides a single weight that will be used to estimate all the variables of interest in the survey. The following set of equations is solved with least squares to minimize the adjustments in the weights.

$$C_j = \sum_{i=1}^{n} w_i x_{ij}$$

Where C_j = Frame Control Total
x_{ij} = Control for farm i, crop j
$w_i \approx 1/\pi_i$

For the June 1997 survey in Minnesota, the weights were calibrated to ensure that estimates of the total control acres for 19 different variables and for the number of farms from the sample exactly equaled corresponding population totals. Table 2 shows the population totals, sample expansion and the ratio of sample over population. The process requires the use of restricted regression as described in Amrhein, Hicks, and Kott (1997) (with the lone restriction that calibrated weights not be smaller than 1.)

Table 2 - Calibration Totals

Item	Control Total (000)	Sample Expansion (000)	Ratio Sample/ Control
Capacity	1670941	1669260	1
Farms	51	52	1.02
Row Crops	13925	14263	1.02
Sunflowers	363	371	1.02
Dry Beans	180	182	1.01
Flax	8	8	1
Small Grains	4039	4255	1.05
Rye	37	29	0.8
Winter Wheat	32	35	1.09
Durum Wheat	12	12	1
Potatoes	79	77	0.97
Barley	668	697	1.04
Corn	6948	7050	1.01
Soybeans	5898	6085	1.03
All Hay	1852	1927	1.04
Alfalfa	1370	1458	1.06
Other Hay	394	387	0.98
Oats	723	741	1.02
Spring Wheat	2477	2621	1.06
Sugar Beets	463	494	1.07

V. Results

The Minnesota test in June 1997 premiered the parallel comparison of sample results from both the old and new methods. A sample drawn using the new design and one drawn using the old design were enumerated at the same time. The samples were selected in such a way that there was an approximate 50% overlap between the two. Farms designated as certainties in the old design were designated as certainties for the new design as well; stratified, simple random samples for the old design were drawn using the CS random numbers from the new design using the fixed sample size method described in Amrhein, Hicks and Kott, 1996. The same questionnaire, enumerators, edit, imputation, and analyses systems were employed for both samples.

Estimated variances were used in the comparison of the two designs. Variances were estimated with the delete-a-group jackknife method (Kott, 1977) treating imputed values as if they were genuine survey values. The finite population correction factor was ignored. Table 3 displays the CV estimates derived from the delete-a-group jackknife for both designs. Estimated CV's from the new design were adjusted to remove the effect of its slightly smaller sample size (i.e., CV's for the new design were multiplied by the square root of the ratio of the old design's sample size to the new design's). The utilization of the new procedure seemed to improve the precision of the estimates for all crop acreage (with the exception of hay) by using the new design. The CV's for stocks from the two designs were very similar.

Table 3: June 1997 Parallel Test 1/

Crop	CV Old Design	CV New Design	Ratio New/ Old
Planted Acreage			
Corn	3.67	1.88	0.51
Soybeans	3.42	2.77	0.81
Spring Wheat	6.91	2.50	0.36
Potatoes	11.79	8.64	0.73
Oats	8.53	8.08	0.95
Dry Beans	11.25	5.91	0.53
Sunflowers	13.36	9.91	0.74
Barley	7.26	6.24	0.86
Alfalfa	4.73	6.96	1.47
Other Hay	19.16	26.45	1.38
Stocks			
Corn	5.20	5.52	1.06
Soybeans	7.13	7.66	1.07
Spring Wheat	10.32	10.26	0.99

1/ Adjusted for sample sizes differences.

One goal of the new design was to target specific sample sizes of rare commodities. Generally this was accomplished, but for the crops of flax and durum wheat there were only 2 and 0 positive reports respectively. This indicates that the control data for these crops is not good enough to find a sufficient number of farmers growing these crops.

VI. Future

Future plans include the parallel comparisons of the two designs in the remaining three quarters and to the expansion of the test to other states.

This innovative design raises many questions:

1. Can the selection probabilities be more efficiently determined?
2. What constraints should be put on the calibrated weights?
3. How should minimum sample sizes be determined?
4. How should adjustments for nonresponse be made?

NASS is actively searching for answers to the questions raised by this new design. For example, Amrhein, Bailey and Fleming (1997) addresses Question 1 and provides a framework for Question 3. Less progress has ben made on Questions 2 and 4. For Question 2, a single constraint that no calibrated weight be less than 1 appears to be effective , although that may change. For Question 4, imputation for non-response is determined using strata definitions from the original design in determining imputation classes. An alternative imputation strategy will be needed when the new design becomes operational.

A related question to how best to adjust for nonresponse is how to measure the precision of a non-response adjusted estimate. If imputation is used to adjust for nonresponse, one possibility is to impute using an analogous methodology within each of the jackknife replicates (presently 15). That idea may prove operationally difficult and in any event needs more research.

VI. References

Amrhein, John F., Hicks, Susan D., and Kott, Phillip S. (1996), "Methods to Control Selection When Sampling from Multiple List Frames,"*ASA Proceedings of the Section on Survey Research Methods*, forthcoming.

Amrhein, John F., Hicks, Susan D., and Kott, Phillip S. (1996), "An Application of a Two-phase Ratio Estimator and the Delete-a-group Jackknife," *ASA Proceedings of the Section on Survey Research Methods*, forthcoming.

Amrhein, John F., Fleming, Charles M. and Bailey, Jeffrey T. (1997), "Determining the Probabilities of Selection in a Multivariate Probability Proportional to Size Sample Design," in *New Directions in Surveys and Censuses: Symposium 97*, Statistics Canada, forthcoming.

Kott, Phillip S. (1997), *Using the Delete-A-Group Variance Estimator in NASS Surveys*, National Agricultural Statistics Service Research Report, forthcoming.