United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research and
Applications
Division

SRB Research Report
Number SRB-91-12

December 1991

EVALUATION OF ESTIMATION OPTIONS FOR THE MONTHLY FARM LABOR SURVEY

Cheryl L. Turner

EVALUATION OF ESTIMATION OPTIONS FOR THE MONTHLY FARM LABOR SURVEY, by Cheryl L. Turner, Research and Applications Division, Ohio Field Research Unit, National Agricultural Statistics Service, United States Department of Agriculture, Washington, D.C. 20250, 1991, NASS Staff Report No. SRB-91-12.

ABSTRACT

The National Agricultural Statistics Service (NASS) currently conducts quarterly Agricultural Labor Surveys (ALS). Eleven states began conducting monthly ALS's in 1991. In this study, a half sample direct expansion and a half sample ratio expansion of the variable "total number of all hired workers" were investigated for possible use in the monthly and seasonal surveys. The data were gathered from eleven states in July and October of 1990. For simulation purposes, July was the quarterly data and the October data composed the monthly data.

Neither the half sample direct expansion nor the half sample ratio expansion outperformed the other as an alternative to the estimate derived from the full sample direct expansion. Two areas of research were recommended to improve both of the half sample expansions. The first area, a weighted estimator, will be explored for its impact on the labor surveys. And the second research area will concentrate on the detection of outliers and their removal.

KEY WORDS

agricultural labor survey, list sampling frame, nonoverlap, direct expansion, ratio expansion

ACKNOWLEDGEMENTS

The author would like to thank Lee Brown for reviewing both the early and final drafts of this paper; I would also like to thank Bill Iwig for his technical advice.

TABLE OF CONTENTS

SUMMARY

The National Agricultural Statistics Service (NASS) currently conducts quarterly Agricultural Labor Surveys (ALS). These ALS are multiple frame surveys consisting of samples selected from both a List Sampling Frame (LSF) and a non-overlap (NOL) portion. The NOL portion consists of a sample of non-overlap Resident Farm Operators (RFO's) from forty percent of the area segments from the June Agricultural Survey (JAS). In 1991, NASS began conducting monthly or seasonal surveys in eleven states. Several sampling plans were tested on the survey variable "total number of all hired workers" for possible use in these monthly/seasonal surveys.

Data were analyzed for the eleven monthly/seasonal states using July and October 1990 ALS data sets. The July data was the quarterly data and, for simulation purposes, the October quarterly data was redefined to be the monthly data set. Estimates of the total number of all hired workers were generated at the state level for the following sampling plans: a half sample direct expansion (half sample DE) and a half sample ratio expansion (half sample RE). Both the half sample DE and the half sample RE were considered as potential alternatives to the current full sample direct expansion (full sample DE).

Two goals were established for the simulated study. In comparing the efficiency of the two half samples, the first goal was to find the "superior" sampling plan. The superior sampling plan would have a consistently smaller mean squared error (mse). The second goal was to evaluate the estimates generated for the total number of all hired workers by the two sampling plans in comparison to the full sample DE for the same survey variable. All three sampling plans should yield approximately the same estimates. While the second goal was achieved and there was no significant difference between the three estimates, the first goal was not achieved. Neither the half sample DE nor the half sample RE distinguished itself as the superior alternative to the full sample DE.

Two areas of further research were recommended for study. A weighted estimator will be explored for its impact on the NOL portion of the labor surveys. This weight will be based on the percentage of the total acres operated which are contained within the enumerated tract. A weighted estimator will effectively increase the pool of farm operations from which a sample will be selected. A second area of research will concentrate on the detection of outliers. Outliers are highly influential observations which can greatly increase the mse within a particular state.

INTRODUCTION

Farm employment estimates have been available since 1909 and farm wage rates since 1866. These estimates have ranged over time from national, to regional, and finally to a combination of regional and state level estimates. In 1975, the Agricultural Labor Survey (ALS), a quarterly estimating program supplanted the previous monthly program. The ALS has remained intact except for a two year period when reductions in program funding necessitated yearly surveys. The ALS is a joint effort between the National Agricultural Statistics Service (NASS), within the United States Department of Agriculture (USDA), and the Department of Labor (DOL).

The population of interest for the ALS is the USDA farm population, which is "all operations that sold or would normally sell at least $1,000 worth of agricultural products the previous year". A sample of farm operators is surveyed during January, April, July, and October of each year to provide estimates of the number of farm workers and of the wage rates paid to the farm workers.

The ALS is a multiple frame survey utilizing a list of medium to large farms as identified on the List Sampling Frame (LSF) and a non-overlap (NOL) portion consisting of a sample of the NOL Resident Farm Operators (RFO's) selected from forty percent of the area segments used in the June Agricultural Survey (JAS). Appendix A contains the LSF and NOL strata definitions. The list is an efficient sampling frame because it is originally stratified on variables relating to the number of hired workers, whereas the area frame is originally stratified solely on the land use. However, the list frame does not completely cover the target population. Therefore, the multiple frame approach is used to combine the efficiency of the list frame with the completeness of the area frame, providing unbiased estimates with adequate precision.

In April 1991, a new labor initiative increased the frequency and scope of the ALS in the major program states. California, Florida, New Mexico, and Texas began conducting monthly agricultural labor surveys. Michigan, New York, North Carolina, Oregon, Pennsylvania, Washington, and Wisconsin were designated as seasonal states. These "seasonal" states will conduct surveys in January and then again in April through October. From these additional surveys, the current estimates will be published for both the total number of all hired workers and the all hired worker wage rates for the four monthly states and the seven seasonal states.

The added frequency of these surveys will greatly increase the respondent burden in the aforementioned states. In an attempt to both reduce this respondent burden and to maintain a "reasonable" coefficient of variation, NASS has conducted a simulated study. The July data was the quarterly data and, for simulation purposes, the October data was redefined to be the monthly data set.

This study utilizes various sampling schemes and expansions in calculating the estimate for the total number of all hired workers. Mean squared errors (mse's) were also generated for the various sampling schemes. The mse's measured how well each sampling scheme estimated the "truth". This paper presents the findings of the simulated study utilizing July and October 1990 Agricultural Labor Survey data. The states included in the study were those eleven monthly and seasonal states - California, Florida, Michigan, New Mexico, New York, North Carolina, Oregon, Pennsylvania, Texas, Washington, and Wisconsin.

OVERVIEW

The simulated study was independently performed on the LSF and the NOL data for each of the eleven monthly and seasonal states. Under each scenario, the July ALS data were the quarterly results (which they actually were) and the October ALS data were treated as the results of a monthly labor survey. The data sets were sampled and expansions were applied to the resulting data sets. Both direct expansions and ratio expansions, and their corresponding mse's were calculated. A direct expansion is an estimate of the population total, where the sampled observations are weighted by the inverse of their probability of selection. A ratio expansion is also an estimate of the population total. A survey-to-base ratio was calculated, where a sample monthly survey record was paired with its corresponding base quarterly survey record. The principle point was that all observations that contribute to the ratio expansion were found in both the monthly survey and the quarterly survey. The resulting ratio was then applied to the base survey's direct expansion.

SAMPLING AND DATA SET CREATION

Sample monthly data sets were created for both the LSF and the NOL data sets from the original October data set. Through sampling, the respondent burden was greatly lessened. But, the cost of this sampling lies in estimates which were less precise or, in other words, an increased mse.

The list sample utilized a replicated sampling scheme. The quarterly (July) data set consisted of two replications, numbered 1 and 2. While the monthly (October) data set consisted of replications 2 and 3. A half sample monthly data set (for both the direct and ratio expansion) was constructed by selecting only replication number 2 from the monthly data set. The full sample monthly data set consisted of data from both replications (and, therefore, all observations) from the monthly data.

As stated earlier, the NOL is composed of the RFO's from forty percent of the JAS area sample. An RFO is a resident farm operator

who lives within the selected segment. A sample of these RFO's was selected for generating the full sample expansions and the same sample was contacted throughout the ALS survey year.

As with the LSF, a half sample monthly data set and a full sample monthly data set of the NOL data were created for calculating both the direct and ratio expansions. The NOL data was originally sorted in state - stratum order, and within each stratum, the data was then sorted by the reporter identification variable. The half sample monthly data set was created by numbering those observations and retaining the even numbered observations. Thus the half sample consisted of one half of the selected RFO's from the monthly data set. Correspondingly, the full sample monthly data set consisted of both the odd and even numbered (all) observations from the monthly data set.

Upon obtaining the monthly sample data sets for the LSF and the NOL samples, "usable data sets" were created for the quarterly data set and for both the half sample and full sample monthly data sets. A "usable data set" consisted of all observations where the response code was neither coded as a refusal nor as an inaccessible, but as a completed interview. Consider the following:

Response Codes
1 = Mail
2 = Telephone Interview
3 = Face to Face Interview
6 = Mail Refusal
7 = Telephone Refusal
8 = Face to Face Refusal
9 = Inaccessible

Thus, all observations containing response codes for mail refusal, telephone refusal, face to face refusal, or an inaccessible (response codes 6, 7, 8, and 9) were excluded from the "usable data set".

Therefore, when applying a direct expansion, the "usable data set" consisted of observations having response codes 1, 2, or 3 in the monthly sample. When calculating a ratio expansion, the "usable data set" consisted of all observations having response codes 1, 2, or 3 in both the monthly sample and the quarterly sample.

ESTIMATION OVERVIEW

After creating the usable data sets for the half sample monthly, full sample monthly, and the quarterly sample, direct expansions and ratio expansions were created for both the LSF and NOL. As mentioned above, the quarterly data were obtained from the usable observations from the July ALS. The monthly data were obtained from the usable October ALS observations. It is important to be

familiar with the sampling procedures because the observations contained within the monthly data set (half or full sample) were entirely dependent upon the sampling procedure used.

For the LSF, the full sample DE (replications 2 and 3) from the monthly data set was considered the "truth". The half sample DE and the half sample RE were two alternatives to the truth. The following LSF estimates were created:

1) Half Sample Direct Expansion -
   The monthly data consisted of the half sample monthly usable data set. The monthly data were then expanded and summed to create state level LSF estimates.

2) Half Sample Ratio Expansion -
   A survey-to-base ratio was created. The monthly data, again consisting of the half sample monthly usable data set, was the survey. The quarterly data was the base. The resulting ratio was a measure of change from the quarterly data to the monthly data. This ratio was then applied to the direct expansion of the quarterly data at the state level to create state level LSF ratio estimates.

3) Full Sample Direct Expansion -
   The monthly data, consisting of the full sample monthly usable data set, were expanded and then summed to create state level estimates. This data set was considered the "truth" and was a base for the comparison of all other LSF alternatives.

Appendix B contains the formulae for the state level LSF direct expansion and ratio expansions.

The following estimates were created for the NOL data sets. As with the LSF, the observations included in the monthly data were dependent upon the sampling scheme used. And, again, the full sample DE from the monthly data set was the "truth", with the half sample DE and the half sample RE being the alternatives.

1) Half Sample Direct Expansion -
   The monthly data were composed of the half sample monthly usable data set. The monthly data were then expanded and summed to create state level NOL estimates.

2) Half Sample Ratio Expansion -
   A survey-to-base ratio was created. As with the LSF, the monthly data, consisting of the half sample monthly usable data set, was the survey and the quarterly data was the base. The resulting ratio was applied to the direct expansion of the quarterly data at the state level to create state level NOL ratio estimates.

3) Full Sample Direct Expansion -
The monthly data consisted of the full sample monthly usable data set. The monthly data was subsequently expanded and then summed to create state level NOL estimates. For the NOL, this data set was considered the "truth" and was a base for the comparison of all other NOL estimates.

Appendix C contains the formulae for the state level NOL direct expansion and ratio expansions.

Both the half sample DE and the half sample RE were compared against each other to determine which was the better alternative estimate to the full sample DE for its respective frame (either LSF or NOL). The basis for the comparison was the mse for the number of All Hired Workers for each alternative. The LSF and NOL estimates were evaluated independently of each other.

CALCULATING THE ESTIMATES

When calculating a direct expansion, the response data of interest (the full and half sample monthly usable data sets) was expanded to the state level. Upon expansion, each observation was then summed to create state level estimates for both the LSF and the NOL.

When creating a ratio expansion, a ratio was based on the comparable observations from the quarterly and monthly usable data sets from each state. All of the observations from the quarterly usable data set (those "comparables" that were used in creating the ratio and those "noncomparables" that were not used in the ratio) were then expanded and summed to the state level and multiplied by the state level ratio. This created an expansion that measured the change from the quarterly data to the monthly data at the state level. The resulting state level ratio, r_s, was:

$$r_s = \begin{cases} \frac{m_s}{q_s}, \text{ if } m_s \geq 0 \text{ and } q_s > 0 \\ 1, \text{ otherwise} \end{cases}$$

where

m_s - *the expanded total of the monthly data for state s*

q_s - *the expanded total of the quarterly data for state s*

r_s - *the ratio for state s*

In the above expression, r_s equaled one when its denominator, q_s, was equal to zero. Therefore, when the expanded quarterly data equaled zero, the resulting state ratio r_s, was set equal to one. This ratio of one essentially equated each corresponding monthly and quarterly data observation within the given state. While the ratio of one (indicating no change from the quarterly to the monthly periods) was a conservative estimate of the measure of change, it still maintained the quality and characteristics of the data.

MEAN SQUARED ERROR

The next step was to compare the efficiency of the two half sample alternatives as estimators of the full sample DE. A simple method for comparing these efficiencies was proposed by Phil Kott in Monthly Labor Indications II: Some NOL Considerations. As indicated previously, the full sample DE for October was considered the "truth" for this study. The objective was to evaluate how well the alternative indications matched this truth value. The mse of each alternative as an estimator of the full sample DE was used for this evaluation. This approach avoids calculating actual design variance estimates based on the complex sample design. The alternative indications for both the LSF and the NOL were:

1) half sample DE, and
2) half sample RE.

Appendix D contains the mse equations for both the LSF and NOL. The mse's were calculated at the state level.

RESULTS

In evaluating the data, a smaller mse for the half sample DE or for the half sample RE indicated which was the better "match" for the full sample DE. Additionally, each estimate represented the total number of all hired workers. Therefore, the full sample DE, the "truth", and each of the half sample alternatives should produce numerically "close" estimates. Appendix E contains the LSF and NOL direct expansion and ratio expansion estimates, and their corresponding mean squared errors for each individual state.

The Fisher Sign Test was performed separately on the LSF and the NOL to determine if there was a significant difference between the mse's for the half sample DE and the half sample RE across all eleven states. Results showed insignificant p-values (p-values of .5000 and .2744 for the LSF and NOL, respectively). These p-values indicate that there was no significant difference between the mse's of the two half sample alternatives for both the LSF and the NOL. Therefore, neither of the half sample mse's distinguished itself as

the superior alternative to match the full sample DE. Table 1 contains the results of the Fisher Sign Test.

The Friedman Rank Sums was used to determine if the estimates from half sample DE and half sample RE were numerically "close" to the estimate from the full sample DE. The test was performed independently on both the LSF and the NOL. Again, the results showed highly insignificant p-values (.976 for the LSF and .732 for the NOL). These p-values indicate that the estimates achieved through the half sample DE and the half sample RE were not significantly different from the estimate of the "truth", the full sample DE. Therefore, each of the half sample expansions sufficiently calculated the full sample DE. Table 2 contains the Friedman Rank Sums results.

TABLE 1: Fisher Sign Test - a comparison of the mean squared errors for the half sample direct expansion and the half sample ratio expansion for both the list sampling frame and the non-overlap.

LSF STATE	MSE HALF SAMPLE DE (000,000)	MSE HALF SAMPLE RE (000,000)	MSE HALF SAMPLE DE - MSE HALF SAMPLE RE (+ or -)
CA	511.13	171.60	+
FL	57.15	102.95	-
MI	32.89	34.02	-
NM	0.81	0.78	+
NY	6.14	6.82	-
NC	4.17	22.00	-
OR	43.49	123.41	+
PA	37.24	3.92	+
TX	35.77	26.63	+
WA	341.58	38.33	+
WI	31.32	31.53	-

Significance Level = .5000

NOL STATE	MSE HALF SAMPLE DE (000,000)	MSE HALF SAMPLE RE (000,000)	MSE HALF SAMPLE DE - MSE HALF SAMPLE RE (+ or -)
CA	27.11	1.51	+
FL	0.00	4.71	-
MI	1.20	1.98	-
NM	0.68	0.19	+
NY	12.69	0.00	+
NC	3.74	3.91	-
OR	4.06	0.54	+
PA	39.36	17.59	+
TX	61.58	50.04	+
WA	9.40	2.02	+
WI	18.22	26.74	-

Significance level = .2744

TABLE 2: Friedman Rank Sums - a comparison of the estimates from the half sample direct expansion and the half sample ratio expansion to the "truth" estimate, the full sample direct expansion, for both the list sampling frame and the non-overlap. The rank of each estimate is in brackets.

LSF STATE	HALF SAMPLE DE (000)	HALF SAMPLE RE (000)	FULL SAMPLE DE (000)
CA	168.25 [1]	192.93 [3]	190.32 [2]
FL	40.59 [2]	40.02 [1]	43.74 [3]
MI	22.55 [2]	28.23 [3]	20.62 [1]
NM	3.29 [1]	5.17 [2]	5.26 [3]
NY	21.01 [1]	25.13 [3]	23.04 [2]
NC	16.30 [2]	15.32 [1]	18.64 [3]
OR	20.20 [3]	17.69 [1]	20.02 [2]
PA	16.51 [2]	19.05 [3]	16.07 [1]
TX	38.28 [2]	34.09 [1]	41.27 [3]
WA	47.50 [3]	46.30 [2]	41.65 [1]
WI	24.76 [2]	26.68 [3]	21.81 [1]

Significance Level = .976

NOL STATE	HALF SAMPLE DE (000)	HALF SAMPLE RE (000)	FULL SAMPLE DE (000)
CA	17.00 [1]	48.95 [3]	36.62 [2]
FL	0.00 [1]	0.41 [2]	1.96 [3]
MI	2.07 [2]	1.75 [1]	3.16 [3]
NM	1.39 [3]	0.00 [1]	0.69 [2]
NY	3.56 [1]	4.83 [3]	4.51 [2]
NC	3.83 [2]	6.83 [3]	2.19 [1]
OR	2.52 [3]	1.66 [2]	1.60 [1]
PA	9.09 [1]	9.54 [2]	11.43 [3]
TX	13.67 [2]	7.79 [1]	13.99 [3]
WA	3.14 [1]	32.85 [3]	8.65 [2]
WI	10.91 [3]	6.15 [1]	10.68 [2]

Significance Level = .732

As Table 2 showed, both the half sample DE and the half sample RE were adequate alternative estimates to the full sample DE. But, based on Table 1, neither of these alternatives distinguished itself as the superior alternative. Neither the half sample DE nor the half sample RE was the "better alternative" in terms of matching the full sample DE. Two techniques are suggested to both improve the accuracy of the estimates and to reduce the mse's: first, improvement within the sample selection processes (especially the NOL); and secondly, the determination of outlier observations.

SAMPLE SELECTION

A sample is "a subset of measurements selected from the population of interest". A half sample implies that one half of the available data was used in creating the estimate. The logical question is "Were the selected LSF and NOL samples sufficient in creating a half sample DE and a half sample RE?"

The sample selection within the LSF is based on a replicated sample design. Although the number of replications drawn has recently changed, the LSF sample design still remains a consistent process. For the surveys conducted from July 1990 through June 1991, four independent replicates were drawn. The quarterly surveys consisted of the selected samples from two replicates, where one replicate was rotated out each quarter. For the monthly surveys conducted prior to July 1991 (May and June 1991), their survey sample contained the same replicates as their preceding quarterly counterpart (April). Beginning in July 1991, eight replicates were drawn (as opposed to four). The quarterly surveys will now consist of the selected samples from four replicates, where two replicates will be rotated out each quarter. These monthly surveys will consist of two replicates (a half sample), with the same replicates being used for both months between the quarterly survey. Therefore, the estimates obtained from the half sample monthly surveys will not be adversely affected due to the consistent, replicated LSF sample design.

As previously mentioned, the NOL portion of the ALS sample was selected from the NOL RFO's contained in forty percent of the JAS area sample. This was done to ease respondent burden between the ALS and the Farm Costs and Returns Survey (FCRS). But, by easing the respondent burden in this manner, a state's quarterly estimate was actually based on the "usable" NOL RFO's contained in a forty percent sample of the JAS area segments. A "usable" NOL RFO does not include the refusals nor the inaccessibles, it is a respondent who gives a valid interview. For both the half sample DE and the half sample RE, the resulting estimates would be based on one half of the "usable" NOL RFO's contained in those same JAS area segments. Therefore, the precision of the estimate within the NOL portion of the ALS (both the quarterly and the monthly surveys)

could be strongly affected by the small "pool" of RFO's from which the sample was selected.

As evidenced in both the LSF and the NOL, a half sample reduces the number of data records, thereby heightening the importance of each individual data record. In the half sample DE and the half sample RE, each record would have two times the impact that the same record had in a full sample DE. A record which was a poor representative of its population, an "outlier", would also have twice its original impact.

OUTLIER OBSERVATIONS

As Hollander and Wolfe defined in their book, Nonparametric Statistical Methods, an outlier is "an observation that is found to lie an abnormally long way from its fellow observations in a series of replicated observations".

Outliers are highly influential observations that affect their estimates. They are present within both the LSF and NOL portions of the ALS. But, an outlier would have very differing effects on the direct and ratio expansions. An observation which was an outlier when creating a direct expansion may loose some of its impact when calculating a ratio expansion. Therefore, an outlier may affect (significantly increase) the mse of a direct expansion while, at the same time, have little affect (no significant increase) on the mse of a ratio expansion.

Recall the state level ratio below.

$$r_s = \begin{cases} \frac{m_s}{q_s}, \text{ if } m_s \geq 0 \text{ and } q_s > 0 \\ 1, \text{ otherwise} \end{cases}$$

where

m_s = *the expanded total of the monthly data for state s*

q_s = *the expanded total of the quarterly data for state s*

r_s = *the ratio for state s*

When considering potential outlier observations (the monthly and quarterly data observations, m_s and q_s, respectively) and their impact on both the half sample DE mse and the half sample RE mse, there were four scenarios.

1) Both m_s and q_s were outliers.
The half sample DE mse would be affected by the presence of the outlier m_s, whether it was a high or a low outlier. The impact on the half sample RE mse would depend on the direction of the outliers. If m_s and q_s were both high or both low (indicating a small magnitude of difference between the two data observations), there would be little affect on the resulting ratio r_s; and therefore, the half sample RE mse would not be affected by these outliers. If m_s were high and q_s were low (or vice versa), r_s would either be very large or very small and the half sample RE mse would be affected.

2) m_s was an outlier, q_s was not
In this instance, the half sample DE mse would again be affected. The large magnitude of difference between the two observations indicates that the half sample RE mse would also be affected.

3) m_s was not an outlier, q_s was an outlier
Since m_s was not an outlier, there was no outlier contained in the monthly half sample, and therefore the half sample DE mse would not be affected. But, as stated above, the half sample RE mse would be affected due to the large magnitude of difference between the two observations.

4) Neither m_s nor q_s were outliers
In this final scenario, neither the half sample DE mse nor the half sample RE mse would be affected since neither the monthly nor the quarterly data observation was an outlier.

To summarize the four scenarios, the half sample DE mse would be affected by an outlier, whereas the half sample RE mse would be affected by a large magnitude of difference - which stemmed from at least one observation being an outlier. When a monthly data observation included in the half sample was an outlier, the half sample DE mse would be affected. When the monthly and\or quarterly data observations included in the half sample were outliers, the half sample RE mse could be affected, depending on the magnitude of difference between the two observations.

Outliers are an added complication to both the half sample DE and the half sample RE. But, there are three possible solutions to the problems presented by outliers. The outlier observation could be the result of a farming operation which was misclassified. If so, updating the control data and reclassifying the farm could possibly place the operation into a strata in which it was not an outlier. Or, for strictly data analysis purposes, there is also the possibility of predetermining the outliers prior to creating the estimates. The outlier observations could be identified and an appropriate robust estimator could then be used. A third possibility does exist. In this scenario, the observations are not

outliers. The "abnormal long way from its fellow observations" is directly related to the seasonal employment of hired workers that is associated with agriculture. Under this scenario there are no outliers and the estimates are representative of the actual data.

RECOMMENDATIONS

Using a half sample DE, half sample RE, and a full sample DE, estimates were generated for the total number of hired workers in each of the eleven monthly and seasonal states. Neither the half sample DE nor the half sample RE proved itself as the superior alternative in matching the full sample DE. Two areas of research were recommended to improve the aforementioned expansions. First, an NOL weighted estimator will be explored for its impact on the labor surveys. The weighted estimator will increase the pool of farm operations and, thereby, enable the sample to be selected from a larger, more representative list of farming operations. In sampling from a larger, more representative pool, it is hoped that fewer outliers would be found. And, the second research area will concentrate on the detection of outliers. The detection of outliers could be a warning sign for a farm misclassification within the strata. By updating the control data and reclassifying the farming operation, the magnitude and impact of the outlier observations could be evaluated.

REFERENCES

[1] Hollander, Myles and Douglas A. Wolfe. Nonparametric Statistical Methods. John Wiley & Sons, New York, NY. 1973.

[2] Kott, Phillip S. "Monthly Labor Indications," U.S. Department of Agriculture, National Agricultural Statistics Service, 1990.

[3] Kott, Phillip S. "Monthly Labor Indications II: Some NOL Considerations," U.S.Department of Agriculture, National Agricultural Statistics Service, 1990.

[4] Kott, Phillip S. "Mathematical Formulae for the 1989 Survey Processing System (SPS) Summary," U.S. Department of Agriculture, National Agricultural Statistics Service, 1990.

[5] U.S. Department of Agriculture (1983): Scope and Methods of the Statistical Reporting Service. Publication No. 1308. Washington, D.C.

[6] U.S. Department of Agriculture, National Agriculture Statistics Service. "Agricultural Labor Survey: Supervising and Editing Manual". June 1990.

APPENDIX A: List Sampling Frame and Non-overlap strata definitions

TABLE 1: List Sampling Frame strata definitions

STRATUM	DESCRIPTION	OPERATIONS INCLUDED
95	Extreme Operators	1. Sheep EO's FVS >= 500,000 2. Poultry EO's FVS >= 500,000 3. Fruit and Veg. farms, FVS >= 500,000 4. Tobacco farms FVS >= 500,000 5. Potato farms FVS >= 500,000 6. Dairy EO's FVS >= 500,000 7. Hog EO's FVS >= 500,000 8. Cattle EO's FVS >= 500,000 9. Other farms FVS >= 500,000
90	Large operators classified on common commodities	1. Sheep EO's, FVS 200,000-499,999 2. Poultry EO's, FVS 200,000-499,999 3. Dairy EO's, FVS 200,000-499,999 4. Hog EO's, FVS 200,000-499,999 5. Cattle EO's, FVS 200,000-499,999 6. All other farms, FVS 200,000-499,999
85	Large operators classified on uncommon commodities	1. Nurseries and greenhouses 2. Fruit and Veg. farms, FVS 100,000-499,999 3. Tobacco farms, FVS 200,000-499,999 4. Potato farms, FVS 200,000-499,999
80	Medium operators classified on common commodities	1. Sheep EO's, FVS <= 199,999 2. Poultry EO's, FVS <= 199,999 3. Dairy EO's, FVS <= 199,999 4. Hog EO's, FVS <= 199,999 5. Cattle EO's, FVS <= 199,999 6. All other farms, FVS 100,000-199,999
75	Medium operators classified on uncommon commodities	1. Fruit and Veg. farms, FVS <= 99,999 2. Tobacco farms, FVS <= 199,999 3. Potato farms, FVS <= 199,999
70	Medium operators classified on estimated sales	1. Farms classified for the Farm Costs and Returns Survey
61-64	Hired worker	1. All farms with BLS control data, stratified on number of hired workers

where,

EO = Extreme Operator
FVS = farm value of sales
BLS = Bureau of Labor Statistics

Special Sampling Situations in California and Florida

California uses a different classification for the farm labor survey because of the availability of extensive control data on the total number of hired workers reported by the State Department of Employment. Their LSF records are stratified exclusively on this hired worker control data.

California and Florida are also the only states that sample lists of agricultural service firms for each survey. In California and Florida, the enumerators interview the agricultural service firms that were reported by the sampled farmers. A multiple frame expansion consisting of both a list portion and an NOL portion is then provided for the agricultural service firms.

TABLE 2: Non-overlap strata definitions

LABOR STRATUM	JAS COMPLETION CODE	DESCRIPTION
10	4, 5	Refusal or inaccessible
9	1, 2, 3	PLF >= 10
8	1, 2, 3	PLF 5 - 9
7	1, 2, 3	PLF 1 - 5
6	1, 2, 3	PLF = 0 (sales code >= 6)
5	1, 2, 3	PLF = 0 (sales code < 6)

where,

PLF = Peak Labor Force

APPENDIX B: List Sampling Frame state level direct expansion and ratio expansion formulae

LSF STATE LEVEL DIRECT EXPANSION FORMULA

$$\hat{Y}_{STATE,LSF,DE} = \sum_{i-1}^{I_m} \frac{pcount_{i,m}}{n_{i,m}} \sum_{j-1}^{J_{i,m}} x_{ij,m}\, LAFX_{ij,m}$$

where

I_m - *the number of list frame strata contained in the monthly usable sample*

$J_{i,m}$ - *the number of sampled tracts within stratum i of the monthly usable sample*

$pcount_{i,m}$ - *the population count within stratum i of the monthly usable sample*

$n_{i,m}$ - *the number of sampled tracts within stratum i of the monthly usable sample*

$x_{ij,m}$ - *the number of paid workers in tract j within stratum i of the monthly usable sample*

$LAFX_{ij,m}$ - *the list adjustment factor for tract j within stratum i of the monthly usable sample*

LSF STATE LEVEL RATIO EXPANSION FORMULA

$$\hat{Y}_{STATE,LSF,RE} = r'_{mq} \sum_{i=1}^{I_q} \frac{pcount_{i,q}}{n_{i,q}} \sum_{j=1}^{J_{i,q}} z_{ij,q}\, LAFZ_{ij,q}$$

where

r'_{mq} - *a state level ratio of respondents who provided data for both the monthly and quarterly usable sampled data sets*

$$= \frac{\textit{monthly sample direct expansion}}{\textit{quarterly sample direct expansion}}$$

$$= \frac{\sum_{i=1}^{I'_{mq}} \frac{pcount'_{i,mq}}{n'_{i,mq}} \sum_{j=1}^{J'_{i,mq}} x'_{ij,mq} LAFX'_{ij,mq}}{\sum_{i=1}^{I'_{mq}} \frac{pcount'_{i,mq}}{n'_{i,mq}} \sum_{j=1}^{J'_{i,mq}} z'_{ij,mq} LAFZ'_{ij,mq}}$$

I'_{mq} - *the number of list frame strata which were contained in both the monthly and quarterly usable samples*

$J'_{i,mq}$ - *the number of sampled tracts from stratum i which were contained in both the monthly and quarterly usable samples*

$pcount'_{i,mq}$ - *the population count from stratum i which was contained in both the monthly and quarterly usable samples*

$n'_{i,mq}$ - *the number of sampled tracts within stratum i which were contained in both the monthly and quarterly usable samples*

$x'_{ij,mq}$ - *the monthly number of paid workers in tract j within stratum i which were contained in the sampled tracts from both the monthly and quarterly usable samples*

$z'_{ij,mq}$ - *the quarterly number of paid workers in tract j within stratum i which were contained in the sampled tracts from both the monthly and quarterly usable samples*

$LAFX'_{ij,mq}$ - *the monthly list adjustment factor for tract j within stratum i from sampled tracts contained in both the monthly and quarterly usable samples*

$LAFZ'_{ij,mq}$ - *the quarterly list adjustment factor for tract j within stratum i from sampled tracts contained in both the monthly and quarterly usable samples*

I_q - *the number of list frame strata contained in the quarterly usable sample*

$J_{i,q}$ - *the number of sampled tracts within stratum i of the quarterly usable sample*

$pcount_{i,q}$ - *the population count within stratum i of the quarterly usable sample*

- $pcount'_{i,mq}$ -- *the population count within stratum i remains constant throughout the survey year*

$n_{i,q}$ - *the number of sampled tracts within stratum i of the quarterly usable sample*

$z_{ij,q}$ - *the number of paid workers in tract j within stratum i of the quarterly usable sample*

$LAFZ_{ij,q}$ - *the list adjustment factor for tract j within stratum i of the quarterly usable sample*

APPENDIX C: Non-overlap state level direct expansion and ratio expansion formula

NOL STATE LEVEL DIRECT EXPANSION FORMULA

$$\hat{Y}_{STATE,NOL,DE} = \sum_{i-1}^{I_m} \frac{t_{i,m}}{n_{i,m}} \sum_{j-1}^{J_{i,m}} x_{ij,m}\, LAFX_{ij,m}\, ADJEFX_{ij,m}$$

where

I_m - *the number of farm labor strata contained in the monthly usable sample*

$J_{i,m}$ - *the number of sampled tracts within stratum i of the monthly usable sample*

$t_{i,m}$ - *the number of tracts within stratum i of the monthly usable sample*

$n_{i,m}$ - *the number of sampled tracts within stratum i of the monthly usable sample*

$x_{ij,m}$ - *the number of paid workers in tract j within stratum i of the monthly usable sample*

$LAFX_{ij,m}$ - *the list adjustment factor for tract j within stratum i of the monthly usable sample*

$ADJEFX_{ij,m}$ - *the adjusted expansion factor for tract j within stratum i of the monthly usable sample*

NOL STATE LEVEL RATIO EXPANSION FORMULA

$$\hat{Y}_{STATE,NOL,RE} = r'_{mq} \sum_{i=1}^{I_q} \frac{t_{i,q}}{n_{i,q}} \sum_{j=1}^{J_{i,q}} z_{ij,q}\, LAFZ_{ij,q}\, ADJEFZ_{ij,q}$$

where

r'_{mq} - *a state level ratio of respondents who provided data for both the monthly and quarterly usable sampled data sets*

$$= \frac{\text{monthly sample direct expansion}}{\text{quarterly sample direct expansion}}$$

$$= \frac{\sum_{i=1}^{I'_{mq}} \frac{t'_{i,mq}}{n'_{i,mq}} \sum_{j=1}^{J'_{i,mq}} x'_{ij,mq} LAFX'_{ij,mq} ADJEFX'_{ij,mq}}{\sum_{i=1}^{I'_{mq}} \frac{t'_{i,mq}}{n'_{i,mq}} \sum_{j=1}^{J'_{i,mq}} z'_{ij,mq} LAFZ'_{ij,mq} ADJEFZ'_{ij,mq}}$$

I'_{mq} - *the number of farm labor strata which were contained in both the monthly and quarterly usable samples*

$J'_{i,mq}$ - *the number of sampled tracts from stratum i which were contained in both the monthly and quarterly usable samples*

$t'_{i,mq}$ - *the number of tracts within stratum i which were contained in both the monthly and quarterly usable samples*

$n'_{i,mq}$ - *the number of sampled tracts within stratum i which were contained in both the monthly and quarterly usable samples*

$x'_{ij,mq}$ - *the monthly number of paid workers in tract j within stratum i which were contained in the sampled tracts from both the monthly and quarterly usable samples*

$z'_{ij,mq}$ - *the quarterly number of paid workers in tract j within stratum i which were contained in the sampled tracts from both the monthly and quarterly usable samples*

$LAFX'_{ij,mq}$ - *the monthly list adjustment factor for tract j within stratum i from sampled tracts contained in both the monthly and quarterly usable samples*

$LAFZ'_{ij,mq}$ - *the quarterly list adjustment factor for tract j within stratum i from sampled tracts contained in both the monthly and quarterly usable samples*

$ADJEFX'_{ij,mq}$ - the monthly adjusted expansion factor for tract j within stratum i from sampled tracts contained in both the monthly and quarterly usable samples

$ADJEFZ'_{ij,mq}$ - the quarterly adjusted expansion factor for tract j within stratum i from sampled tracts contained in both the monthly and quarterly usable samples

I_q - the number of farm labor strata contained in the quarterly usable sample

$J_{i,q}$ - the number of sampled tracts from stratum i of the quarterly usable sample

$t_{i,q}$ - the number of tracts within stratum i of the quarterly usable sample

$n_{i,q}$ - the number of sampled tracts within stratum i of the quarterly usable sample

$z_{ij,q}$ - the number of paid workers in tract j within stratum i of the quarterly usable sample

$LAFZ_{ij,q}$ - the list adjustment factor for tract j within stratum i of the quarterly usable sample

$ADJEFZ_{ij,q}$ - the adjusted expansion factor for tract j within stratum i of the quarterly usable sample

APPENDIX D: LSF and NOL state level Mean Squared Error direct expansion and ratio expansion equations

LSF AND NOL STATE LEVEL DIRECT EXPANSION MEAN SQUARED ERROR FORMULA

$$MSE_{STATE,DE} = \sum_{i-1}^{I_m} n_{i,m}\, s_{i,m}^{2}$$

where

I_m - *the number of strata (list frame or land use) contained in the monthly usable sample*

$n_{i,m}$ - *the number of sampled tracts within stratum i of the monthly usable sample*

$$s_{i,m}^{2} = \frac{\sum_{j-1}^{n_{i,m}} x_{ij,m}^{2} - \frac{\left(\sum_{j-1}^{n_{i,m}} x_{ij,m}\right)^{2}}{n_{i,m}}}{n_{i,m} - 1}$$

$x_{ij,m}$ - *the expanded number of paid workers in tract j within stratum i of the monthly usable sample*

LSF AND NOL STATE LEVEL RATIO EXPANSION MEAN SQUARED ERROR FORMULA

$$MSE_{STATE,RE} = \sum_{i-1}^{I_{mq}} n_{i,mq} \; s_{i,mq}^{2}$$

where,

I_{mq} - *the number of strata (list frame or land use) which were contained in both the monthly and quarterly usable samples*

$n_{i,mq}$ - *the number of sampled tracts within stratum i which were contained in both the monthly and quarterly usable samples*

$$s_{i,mq}^{2} = \frac{\sum_{j-1}^{n_{i,mq}} e_{ij,mq}^{2} - \frac{(\sum_{j-1}^{n_{i,mq}} e_{ij,mq})^{2}}{n_{i,mq}}}{n_{i,mq} - 1}$$

$e_{ij,mq}$ - *a measure of change from the monthly sample to the quarterly sample of the expanded number of paid workers in tract j within stratum i which were contained in sampled tracts from both the monthly and quarterly usable samples*

$$= x_{ij,mq} - (\frac{\tilde{T}_M}{\tilde{T}_Q}) \; z_{ij,mq}$$

$x_{ij,mq}$ - *the monthly expanded number of paid workers in tract j within stratum i which were contained in the sampled tracts from both the monthly and quarterly usable samples*

$z_{ij,mq}$ - *the quarterly expanded number of paid workers in tract j within stratum i which were contained in the sampled tracts from both the monthly and quarterly usable samples*

$\tilde{T}_M$ - *a monthly direct expansion estimate of the number of paid workers within stratum i which was based upon tracts contained in both the monthly and quarterly usable samples*

$\tilde{T}_Q$ - *a quarterly direct expansion estimate of the number of paid workers within stratum i which was based upon tracts contained in both the monthly and quarterly usable samples*

APPENDIX E: State level estimates and mean squared errors

TABLE 1: List Sampling Frame results

STATE	HALF DE (000)	MSE HALF DE (000,000)	HALF RE (000)	MSE HALF RE (000,000)	"TRUTH" FULL DE (000)
CA	168.25	511.13	192.93	171.60	190.32
FL	40.59	57.15	40.02	102.95	43.74
MI	22.55	32.89	28.23	34.02	20.62
NM	3.29	0.81	5.17	0.78	5.26
NY	21.01	6.14	25.13	6.82	23.04
NC	16.30	4.17	15.32	22.00	18.64
OR	20.20	43.49	17.69	123.41	20.02
PA	16.51	37.24	19.05	3.92	16.07
TX	38.28	35.77	34.09	26.63	41.27
WA	47.50	341.58	46.30	38.33	41.65
WI	24.76	31.32	26.68	31.53	21.81

TABLE 2: Non-overlap results

STATE	HALF DE (000)	MSE HALF DE (000,000)	HALF RE (000)	MSE HALF RE (000,000)	"TRUTH" FULL DE (000)
CA	17.00	27.11	48.95	1,509.45	36.62
FL	0.00	0.00	0.41	4.71	1.96
MI	2.07	1.20	1.75	1.98	3.16
NM	1.39	0.68	0.00	0.19	0.69
NY	3.56	12.69	4.83	0.00	4.51
NC	3.83	3.74	6.83	3.91	2.19
OR	2.57	4.06	1.66	0.54	1.60
PA	9.08	39.36	9.54	17.59	11.43
TX	13.67	61.58	7.79	50.04	13.99
WA	3.14	9.40	32.85	2.02	8.65
WI	10.91	18.22	6.15	26.74	10.68

U.S. Government Printing Office [illegible] 311-355/60008