

EVALUATION OF ESTIMATION OPTIONS FOR THE MONTHLY FARM LABOR SURVEY

Cheryl L. Turner, USDA/OASS
200 N. High, Room 608, Columbus, OH 43215

KEY WORDS

agricultural labor survey, list sampling frame, non-overlap

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). 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.

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 (DE) and ratio expansions (RE), and their corresponding MSE's were calculated.

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) of 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

ESTIMATION OVERVIEW

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 all (both the odd and even numbered) 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

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.

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 both LSF and the NOL, the full sample DE from the monthly data set was considered the "truth". The half sample DE and the half sample RE were two alternatives to the truth. Both LSF and NOL estimates were created for each of the following:

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 and NOL estimates.

2) Half Sample Ratio Expansion: A survey-to-base ratio was created. The monthly data, 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 and NOL 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 and NOL alternatives.

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) were 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, the 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.

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.

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.

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.

Therefore, 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; and secondly, the determination of outlier observations.

RECOMMENDATIONS

Using a half sample DE, a 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. 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.