United States Department of Agriculture

National Agricultural Statistics Service

Research Division

RD Research Report Number RD-97-06

October 1997

An Evaluation of List-Only, Reweighted, and Other Estimators for U.S. Agricultural Labor Surveys

Charles R. Perry
Raj S. Chhikara
Floyd M. Spears
Susan Cowles
William C. Iwig

AN EVALUATION OF LIST-ONLY, REWEIGHTED, AND OTHER ESTIMATORS FOR U.S. AGRICULTURAL LABOR SURVEYS by Charles R. Perry, Raj S. Chhikara[1], Floyd M. Spears[2], Susan Cowles and William C. Iwig. Sampling and Estimation Research Section, Research Division, National Agricultural Statistics Service, United States Department of Agriculture, Washington DC 20250-2000, July 1997, Report No. RD-97-06, October 1997.

ABSTRACT

This report describes a comprehensive study of a number of list-based estimators that would minimize area frame sampling for incompleteness of the list frame when estimating hired, self-employed, and unpaid workers from the Quarterly Agricultural Labor Surveys (QLS). The study was based on data from sixteen quarterly labor surveys during 1992-96. The evaluation compares the currently used direct expansion (DE) estimator and seven alternative estimators which can be grouped into three categories according to the level of use that is made of the not on the list (NOL) samples: (1) a difference estimator and a reweighted expansion estimator, along with the DE, that require NOL samples each quarter, (2) two difference estimators, one based on the DE and the other based on a regression-type estimator, along with a ratio estimator, that require only July NOL samples, and (3) a post-stratified estimator and a regression-type estimator, that require no NOL samples.

Using a jackknife procedure, the bias and variance are estimated for each estimator. In the case of the DE estimator, the currently used variance estimator is also contrasted with an unbiased variance estimator. None of the list-based or list-only estimators is uniformly superior, although some of them compared favorably with the DE in estimating the items for hired workers at the U.S. level. Overall, the ratio estimator has an advantage over the other alternative estimators, since it is easy to implement and it performs as well as any of the list-based and list-only estimators.

KEY WORDS

Direct Expansion; Post-stratification; Ratio and Regression-Type Estimators; Jackknife; Variance; Bias.

ACKNOWLEDGEMENTS

The authors would like to thank Ron Bosecker and George Hanuschak for their continued support and guidance throughout the work on this project.

---

[1] Raj S. Chhikara is Professor, Division of Computing and Mathematics, University of Houston-Clear Lake.

[2] Floyd M. Spears is Assistant Professor, Division of Computing and Mathematics, University of Houston-Clear Lake.

Contents

List of Figures

List of Tables

SUMMARY

This report summarizes a comprehensive evaluation of eight estimators for measuring not on the list (NOL) incompleteness in hired, self-employed, and unpaid workers from the Quarterly Agricultural Labor Survey (QLS). The motivation for looking at five of these estimators was to minimize the respondent burden associated with contacting NOL area frame respondents quarterly for their labor information. For comparison, the eight estimators were grouped into three categories according to the use made of the NOL samples.

- **Quarterly labor NOL samples used:** This category consisted of the currently used DE, a reweighted estimator similar to the DE, and a regression-type estimator, called the difference estimator, that makes use of auxiliary information from both the QLS and the June Agricultural Survey (JAS).
- **July labor NOL samples used:** This category consisted of a ratio estimator and two list-based estimators that update the July NOL estimates with the difference between the NOL regression estimates for the current and July quarters. The ratio estimator was obtained by multiplying the quarterly labor survey list estimate by the ratio of the multiple frame estimate to the list estimate for the July quarter. One list-based estimator, the list-based direct expansion, was obtained by adding the difference to the July NOL DE estimate. The other list-based estimator, the list-based difference, was obtained by adding the difference to the July NOL difference estimate.
- **No NOL samples used:** This category consisted of a post-stratified estimator and a regression-type estimator called the predicted NOL. The post-stratified estimator used post-stratum weights from the JAS and post-stratum means from the QLS. The predicted NOL estimator used the regression fits from the QLS list frame samples to predict for all NOL tracts from the JAS area samples.

For each of the eight estimators, sixteen quarterly estimates were computed using the 1992-96 QLS data. A set of estimates was derived for each of the items:

- Total number of hired, self-employed, and unpaid workers.
- Average hours worked in a week for hired, self-employed, and unpaid workers.
- Weekly wage rates for hired workers.

Each set of estimates was initially compared with the official quarterly statistics, called Board estimates. The relative mean deviation from the Board (R-MD) and the relative root mean square deviation from the Board (R-RMSD) were computed for each item at the U.S. level. Overall, the DE was slightly higher than the Board and its R-RMSD was, at most, 5%. Some, if not all, of the differences between the DE and the Board may be attributed to two factors: 1) the use of archived files instead of the data available at the time the

Board estimates were set, and 2) the fact that outliers inherent in the survey data are often re-evaluated when setting official statistics.

The performance of the reweighted estimator with respect to the Board was similar to that of the DE. The difference estimator was inferior to the DE in estimating total number of hired, self-employed, and unpaid workers. None of the estimators in the other two categories performed as well as the DE and reweighted estimators. However, the ratio estimator performed best among the estimators in the other categories.

A similar evaluation was made for each alternative estimator with respect to its R-MD and R-RMSD from the DE. There was very little difference between the reweighted estimator and the DE. Overall, the difference and ratio estimators were similar, for example, each had approximately 4% R-RMSD for the total hired workers. The other estimators had higher R-MD and R-RMSD, particularly for the weekly hours.

A comparison was made between the current variance estimator for the DE and a proposed unbiased variance estimator. This comparison showed that the current variance estimator was robust and reliable. Overall, in the case of hired workers, it was only 2% higher than the unbiased variance estimator. However, the unbiased variance estimator was found to be completely unreliable in the case of self-employed workers.

Each estimator was evaluated for both bias and variance using a jackknife procedure. Relative to the DE, the reweighted estimator was unbiased in every case. All other estimators did not exhibit any significant bias relative to the DE for the total hired workers. But, most were slightly, though insignificantly, biased in estimating self-employed workers. The average CVs for the reweighted, difference, ratio and predicted NOL that were computed with the jackknife procedure were similar. However, the CV of the predicted NOL estimator was the smallest overall.

Jackknife relative bias and variance estimates were computed for each estimator. Using 95% confidence intervals for the jackknife relative bias, none of the estimators showed any significant bias. However, with one exception, all alternative estimators had relatively large variance in comparison to the DE. The exception was the predicted NOL which had a smaller variance than the DE in all but the case of unpaid workers. It should be noted that in the case of self-employed workers the variance estimates for the reweighted estimator were as much as 21 percent higher than for the DE.

To summarize, none of the alternative estimators matched the performance of the DE. In the category of "Quarterly labor NOL samples used", the reweighted estimator performed better than the difference estimator. In the category of "July labor NOL samples used", the ratio estimator was the most consistent. In the category of "No NOL samples used", the predicted NOL estimator performed better than the post-stratified estimator. Overall, the ratio estimator has an advantage over the other list-based and list-only estimators, since it is easy to implement and it performs as well any of the others.

INTRODUCTION

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) uses a multiple frame sampling and estimation methodology where the area frame sample is used to compensate for the lack of coverage by the list frame. The area frame samples are separated into two domains, those that are on the list and those that are not on the list (NOL). The NOL sample size is considerably smaller than the list frame sample size. Vogel (1995) discusses the rationale of the multiple frame sampling, outlines the estimation approach and describes the difficulties involved in its implementation. The cost of data collection for area samples is relatively high, it adds to the respondent burden, and the precision of estimates based on NOL area samples is low.

Initially, Vogel, among others, suggested that NASS consider development of alternative approaches that would either minimize or eliminate the use of area frame samples for estimation of a labor characteristic. One basic idea was to develop a list-only approach to estimation using list frame samples from the Quarterly Agricultural Labor Survey (QLS) data and the annual June Agricultural Survey (JAS) data. The initial approach involved post-stratification of the sample data for both the current survey period and the annual JAS and development of certain post-stratified estimators. The post-stratification was based on farm types, farm value of sales, and the peak number of workers.

An evaluation of this approach was made using the monthly agricultural labor survey data from California and Florida for the survey periods between April 1991 and November 1992. The results of this study, which are described in Rumburg, et al (1993), show that the post-stratified estimates often differed substantially from the official Board estimates and thus are not reliable. Another study followed which involved the use of a regression type estimator constructed using auxiliary variables. The approach was based on post-stratification by farm type and the use of the other auxiliary variables used as regressors. Under this approach, the list samples are used to obtain the least-square fits. The resulting equations are then used to predict the NOL totals using the auxiliary information available from the JAS.

Certain evaluations were made using this approach for estimation of agricultural labor characteristics in California and Florida as described in Chhikara, et al. (1995). It was shown that within post-strata each hired worker characteristic (total number, weekly hours and wage rates) was fairly well correlated with the farm value of sales and the peak number of workers and that the estimates obtained from the regression type estimator compared favorably with the current direct expansion estimates. Another study followed which evaluated two regression type estimators and two post-stratified estimators as described in Spears, et al. (1996) which used QLS data from July 1992 until April 1995 for the entire U.S. Overall, none of the alternative estimators performed as well as the direct expansion estimator on a consistent basis; however, in some cases the performance of the alternative estimators was comparable.

In the present study, three list-based estimators are developed. Two of these estimators, the list-based direct expansion and the list-based difference, are based on the regression approach. The other list-based estimator is obtained by multiplying an item's list estimate for a survey period by the ratio of the item's July multiple frame estimate

to its July list estimate. Also considered are two estimators that require no NOL samples. One estimator, the predicted NOL, uses the predicted values for the NOL from regression fits obtained from the labor list samples to estimate the NOL. The other estimator evaluated is a post-stratified list-only estimator. The present study also included the currently used direct expansion estimator and another estimator similar to it, called the reweighted expansion estimator.

First, the agricultural labor survey data and the relevant auxiliary variables are briefly described. This is followed by a description of the various estimators that were evaluated in this study. Next, described are the numerical evaluations of these estimators made using the 1992-93, 1993-94, 1994-95 and 1995-96 quarterly survey data for the number of workers and the average hours worked per week for hired, self-employed and unpaid workers as well as the hourly wage rates for hired workers. All estimators are compared with the Board as well as the currently used direct expansion estimator (DE) which requires the use of quarterly NOL sample data. This is followed by an evaluation of the variance estimation for the DE. Lastly, the bias and the variance of all estimators are evaluated using the jackknife procedure.

LABOR SURVEY DATA

Variables and Models

The quarterly labor surveys are conducted to estimate various characteristics for the three types of farm labor: hired workers, self-employed workers, and unpaid workers. The survey response variables for all three labor types are total number of workers employed and total hours worked for a specified week, and for hired workers, the total wages paid for a specified week. For hired, self-employed, and unpaid workers, the variables of interest are (1) total workers for a specified week (y_1, y_4, y_6), (2) average weekly hours per worker for a specified week (y_2, y_5, y_7) and (3) the average hourly wage rate per hired worker for a specified week (y_3). The auxiliary information from the JAS and the QLS that is used in the development of regression type estimators for the NOL includes: Farm Value of Sales (x_1), Peak Number of Workers (x_2), Farm Type (x_3), and Number of Partners (x_4). These auxiliary variables were examined for use in post-stratifing the sample data or in obtaining the regression fits. Only the variables which showed useful association or linear relationship with a response variable were used.

The farm type, a categorical variable, was found to effect an efficient post-stratification of sample data. On the other hand, the peak number of workers was found to have the highest correlation with the hired number of workers. The farm value of sales, also a categorical variable which has eight or nine sale value categories, was converted to a numeric variable to be used as a regressor. However, no significant correlation was found between farm value of sales and any of the response variables after taking into account the peak number of workers. Only the number of partners had any significant correlation with the number of self-employed workers. None of the auxiliary variables were correlated with the number of unpaid workers.

Scatter plots and correlation coefficients are given in Appendix A for a few representative post-strata and regions. The scatter plots for the hired number of workers versus the peak number of workers show that a non-intercept linear regression model that uses peak number of workers as the regressor

variable is appropriate for the hired workers for each of the variables y_1, y_2 and y_3. Similarly a non-intercept linear regression model that uses the number of partners as the regressor is appropriate for the self-employed workers for variables y_4 and y_5. A fixed mean model is the only appropriate model for the unpaid workers for the variables y_6 and y_7.

Post-Stratification

The post-stratification by farm-type often resulted in post-strata that did not all contain sufficient data to reliably estimate their regression function. This problem was remedied by collapsing farm type post-strata so that there were at least 15 sample observations in each post-stratum. The collapsing procedure involved initial computation of regression coefficients for all post-strata based on annual JAS (historical) data. If the smallest post-strata had less than 15 sample observations, it was collapsed with the closest post-strata measured in terms of the distance between the regression coefficients. The regression coefficients were then updated and the procedure was repeated until the smallest post-strata had at least 15 sample observations.

Further evaluations of the estimators studied here were made by considering post-stratification with a minimum of 30 and 60 sample observations per post-stratum. When the results were compared to the case of a minimum of 15 observations per post-stratum, no significant differences were found in the evaluation results of these estimators.

ESTIMATORS

Currently, NASS employs a multiple frame estimation methodology that combines separate, independently computed, direct expansion estimates of the list and NOL components into an estimate of the state or regional total. The direct expansion estimator of the list component is formulated as follows:

$$ {}_{\text{list}}\hat{Y}_{\text{de}} = \sum_{l=1}^{L} \sum_{i \in s_l} w_i y_i, \qquad (1) $$

where w_i is the expansion factor and y_i is the observed value for the ith sample unit in list stratum l; and s_l denotes the set of sample units in stratum l. Since the list frame covers at least 70 percent of the population, this estimator is based on a sample which is large enough to ensure that it is efficient. A detailed description of both the list and NOL components of the multiple frame direct expansion estimator is given in Kott (1991).

The present study focuses on the development of a more efficient, yet cost effective, estimator of the NOL component than is currently available. The eight estimators developed and evaluated can be grouped into three categories with respect to use of the NOL.

Quarterly Labor NOL Samples Used in Estimation

The estimators described in this subsection are for the NOL component of the current quarter. The total is estimated by adding this to the direct expansion estimate of the list component for the current quarter. Thus, the NOL sample is required in each quarterly labor survey period.

Direct Expansion (DE): The NOL component estimator is,

$$ {}_{\text{nol}}\hat{Y}_{\text{de}} = \sum_{h=1}^{H} \sum_{i \in s_h} w_i y_i \qquad (2) $$

where w_i is the expansion factor and y_i is the observed value for the ith sample unit in area stratum h; and s_h denotes the set of its sample units.

Reweighted Expansion (ReWt): The NOL component estimator is,

$$ {}_{\text{nol}}\hat{Y}_{\text{rewt}} = \sum_{h=1}^{H} \left(\sum_{i \in s_h} w_i \left(\frac{\sum_{j \in S_h} w_i}{\sum_{j \in s_h} w_i} \right) a_i y_i \right) \tag{3} $$

where w_i is the first phase expansion factor for the ith sample unit of the second phase farm labor stratum h, y_i is the observed value for this sample unit, and a_i is the adjustment due to tract to farm ratio, nonresponse, and data adjustment as a result of farm partnership, etc. S_h is the set of farm operators from the first phase units and s_h is the set of sampled units in stratum h.

Note: If, on the other hand, the tract to farm ratio is considered as part of the first phase expansion, then it leads to slightly different values for the w_i and the a_i, and thus, to another NOL component estimate. The reweighted estimates were computed both ways and there was very little difference found between the results obtained using the two methods of computation.

Difference Estimator (Diff): This regression-type estimator requires the post-stratification of the list and NOL samples by farm type followed by a regression fit of the response y onto an auxiliary variable x using only the current *list* sample data. If the best-fit equation is

$$ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, $$

then the predicted response for each NOL sample in a post-stratum is obtained by

$$ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \tag{4} $$

where x_i is the auxiliary variable value for the ith NOL sample. Once this prediction is carried out for all NOL samples from the JAS in each post-stratum, the following two estimates of the NOL component are computable:

$$ {}_{\text{nol}}\hat{Y}_{\text{p,100}} = \sum_{i=1}^{n_J} \hat{y}_i e_{Ji}, \tag{5} $$

and

$$ {}_{\text{nol}}\hat{Y}_{\text{p,40}} = \sum_{i=1}^{n_L} \hat{y}_i e_{Li}, \tag{6} $$

where 100 in the subscript denotes all (100%) of the June NOL tracts, 40 in the subscript denotes NOL tracts in the 40% sampled for Labor, e_{Ji} and e_{Li} are the JAS and QLS expansion factors for NOL sample i.

The NOL component difference estimator is

$$ {}_{\text{nol}}\hat{Y}_{\text{di}} = {}_{\text{nol}}\hat{Y}_{\text{p,100}} + \hat{D}, \tag{7} $$

where ${}_{\text{nol}}\hat{Y}_{\text{p,100}}$ is the predicted NOL using the auxiliary data for the 100% NOL sample from the JAS as shown in Equation 5, and

$$ \hat{D} = {}_{\text{nol}}\hat{Y}_{\text{p,40}} - {}_{\text{nol}}\hat{Y}_{\text{de}} $$

which is the difference between the predicted and actual estimates computed from the NOL samples for the current quarter.

July Labor NOL Samples Used in Estimation

The estimators described in this subsection are for the NOL component of the current quarter. In each case, the current quarter is estimated by adjusting the July NOL estimate. Thus, no NOL samples are required in the current quarter. The total is estimated by adding the direct expansion estimate of the list component for the current quarter to the NOL component estimate.

List-Based Direct Expansion (LBDE): The NOL component estimator is,

$$_{\text{nol}}\hat{Y}_{\text{lbde}} = {}_{\text{nol}}\hat{Y}_{\text{de,J}} + ({}_{\text{nol}}\hat{Y}_{\text{p,L}} - {}_{\text{nol}}\hat{Y}_{\text{p,J}}) \quad (8)$$

where $_{\text{nol}}\hat{Y}_{\text{de,J}}$ is the direct expansion for July of the current year and $(_{\text{nol}}\hat{Y}_{\text{p,L}}\text{-}_{\text{nol}}\hat{Y}_{\text{p,J}})$ is the difference between the predicted NOL for the current quarter and July of the current year.

List-Based Difference (LBDiff): The NOL component estimator is,

$$_{\text{nol}}\hat{Y}_{\text{lbdi}} = {}_{\text{nol}}\hat{Y}_{\text{di,J}} + ({}_{\text{nol}}\hat{Y}_{\text{p,L}} - {}_{\text{nol}}\hat{Y}_{\text{p,J}}) \quad (9)$$

where $_{\text{nol}}\hat{Y}_{\text{di,J}}$ is the difference estimate for July of the current year and $(_{\text{nol}}\hat{Y}_{\text{p,L}}\text{-}_{\text{nol}}\hat{Y}_{\text{p,J}})$ is the difference between the predicted NOL for the current quarter and July of the current year.

Ratio: The NOL component estimator is,

$$_{\text{nol}}\hat{Y}_{\text{ratio}} = {}_{\text{list}}\hat{Y}_{\text{de}} \left(\frac{{}_{\text{nol}}\hat{Y}_{\text{de,J}}}{{}_{\text{list}}\hat{Y}_{\text{de,J}}} \right) \quad (10)$$

where $_{\text{list}}\hat{Y}_{\text{de}}$is the current direct expansion estimate of the list and $_{\text{nol}}\hat{Y}_{\text{de,J}}$ and $_{\text{list}}\hat{Y}_{\text{de,J}}$ are the July direct expansion estimates of the NOL and list, respectively.

No NOL Samples Used in Estimation

The estimators described in this section do not require NOL samples from any quarterly survey and are referred to as "list-only" estimators.

Predicted NOL (PNOL): The NOL component estimator is,

$$_{\text{nol}}\hat{Y}_{\text{p}} = {}_{\text{nol}}\hat{Y}_{\text{p,100}} \quad (11)$$

where $_{\text{nol}}\hat{Y}_{\text{p,100}}$, the predicted NOL uses regression coefficients from the current list survey and the 100% NOL samples from the JAS and is defined in Equation 5. An estimator for total is obtained by adding the current list estimate to the predicted NOL estimate for the current period.

The total is estimated by adding the direct expansion estimate of the list component to the predicted NOL estimate for the current quarter.

Post-Stratified (PSAF): The post-stratified area frame estimator does not use the current list estimate. The total is estimated directly as given by:

$$\hat{Y}_{\text{psaf}} = \sum_{k=1}^{K} \hat{N}_{.k} \hat{\bar{y}}_k \quad (12)$$

where $\hat{N}_{.k}$ is the estimated size for post-stratum k using the JAS data and

$$\hat{\bar{y}}_k = \frac{\sum_{i \in U_k} w_i y_i}{\sum_{i \in U_k} w_i}$$

where w_i is the weight of the ith labor list sample unit that falls in post-stratum k and y_i is its observed value, and U_k is the set of labor list sample units in that post-stratum.

EMPIRICAL EVALUATIONS

The estimators discussed in the previous section were evaluated using the 1992-93, 93-94, 94-95 and 95-96 Quarterly Agricultural

Labor Surveys from states in each of the 17 agricultural regions in the United States. Estimates of the labor items of interest (total number of workers and hours worked per week for the hired, self-employed and unpaid workers, as well as the wage rates for the hired workers) were computed first at the regional level and then aggregated to the national level. These estimates were compared with the corresponding Board statistics. The alternative estimators were also compared with the direct expansion estimator since it is the current procedure used to obtain both the list and NOL estimates.

The relative mean deviation (R-MD) and the relative root mean squared deviation (R-RMSD) were used for the performance criterion. The relative mean deviation from the DE given by,

$$\text{R-MD}(\hat{Y}) = \frac{\sum_{i=1}^{16}(\hat{Y}_i - \hat{Y}_{DE,i})}{\sum_{i=1}^{16}\hat{Y}_{DE,i}}, \qquad (13)$$

measures the average deviation of an estimator ($\hat{Y}$) from the corresponding DE ($\hat{Y}_{\text{de}}$) relative to the DE using the 16 quarterly survey estimates that were computed. The relative root mean squared deviation from the DE given by,

$$\text{R-RMSD}(\hat{Y}) = \frac{\sqrt{\frac{1}{16}\sum_{i=1}^{16}(\hat{Y}_i - \hat{Y}_{DE,i})^2}}{\frac{1}{16}\sum_{i=1}^{16}\hat{Y}_{DE,i}}, \qquad (14)$$

measures the relative variability of an estimator from the corresponding DE relative to the DE using the 16 quarterly survey estimates.

The R-MD and R-RMSD were computed at the state, regional and US level for each labor item for each of the alternative estimators. Similarly, in order to make comparisons with the Board estimates, the R-MD and R-RMSD from the Board were also computed for each of the estimators, including the DE.

Comparison to Board at US Level

Listed in Table 1 on page 7 are the R-MD and R-RMSD for each estimator relative to the official Board statistics at the US level. These results indicate that the DE is in general slightly higher than the Board and that the relative root mean square deviation of the DE from the Board ranges from 0.7% in the case of average weekly hours of hired workers to 5% in the case of total number of hired workers as well as unpaid workers. It should be noted that sample data used in all analyses reported here are from the archived files, which do not always correspond exactly to the files used in computing the estimates that are utilized in deriving the official statistics. This, along with the fact that the outliers inherent in survey data are discounted by the board in setting the official statistics, probably accounts for some, if not most, of the deviation of the official statistics from the DE estimates computed in this study.

The following conclusions can be derived based on the R-RMD and R-RMSD from the Board as seen in Table 1 on page 7.

- The reweighted estimator does consistently well in estimating each labor item, with R-MD and R-RMSD values similar to those of the DE.
- The ratio estimator does well, with R-MD and R-RMSD values slightly higher than those of the DE, in estimating labor items for hired workers. However, it does not perform as well as

Table 1: **R-MD and R-RMSD Relative to Board (US Level)**

Item Estimated	Estimator DE	Diff	LBDE	LBDiff	PNOL	PSAF	Ratio	ReWt
				R-MD(%)				
				Hired				
Total	3.2	6.1	4.1	6.8	6.8	2.9	5.3	3.3
Weekly Hours	-0.3	-0.3	-0.9	-0.5	0.8	2.3	-0.6	-0.3
Wage Rate	1.7	1.6	3.0	2.4	0.6	2.0	2.0	1.7
				Self-Employed				
Total	1.9	9.3	6.2	8.9	13.6	-0.2	4.5	1.9
Weekly Hours	-1.1	-0.4	-6.4	-5.5	17.1	18.2	-1.9	-1.1
				Unpaid				
Total	-0.8	6.3	0.2	2.6	28.1	10.0	5.0	-0.7
Weekly Hours	0.2	0.4	-5.4	-4.4	3.4	5.5	-0.1	0.2
				R-RMSD(%)				
				Hired				
Total	5.0	7.2	7.4	8.7	8.2	6.6	7.1	5.0
Weekly Hours	0.7	0.8	1.6	1.0	1.1	2.7	1.0	0.8
Wage Rate	3.1	3.0	4.4	3.8	2.7	3.9	3.6	3.1
				Self-Employed				
Total	3.3	9.7	6.9	9.3	14.3	8.3	5.0	3.2
Weekly Hours	1.4	1.2	8.9	8.0	18.3	19.1	2.8	1.4
				Unpaid				
Total	5.0	8.0	9.6	9.9	29.7	18.5	7.4	4.8
Weekly Hours	1.0	1.0	8.9	7.7	4.2	5.9	1.9	1.0

the DE in estimating for self-employed and unpaid workers.

- The difference estimator does consistently well for weekly hours for each type of farm labor, but not for the total number of workers.
- The predicted NOL is inconsistent. It is as good as any other estimator in the case of average weekly hours and wage rates of hired workers. However, it performs very poorly for some of the other items estimated.
- The LBDE, LBDiff and post-stratified estimators show mostly poor performance compared to the others.

Comparison to Direct Expansion

The R-MD and R-RMSD from the DE were computed for the alternative estimators con-

sidered in this study. The results are given at the US level in Table 2 on page 9 and at the regional level in Figures B1-B14 of Appendix B. The following conclusions can be drawn from these results:

- The reweighted estimates are almost the same as the DE. There are some exceptions to this at the regional level; the estimates of wage rates for hired workers and weekly hours for self-employed workers in several regions, including California, have R-RMSD of approximately 10 percent.
- The difference estimates of total workers are on the average higher than DE estimates, with R-RMSD of 3.6 percent for the hired and more than 8 percent for the self-employed and unpaid workers. Estimates for the two estimators are similar for weekly hours and wage rates, with R-RMSD less than 1 percent.
- For the three list-based estimators (LBDE, LBDiff and ratio), estimates of total workers are on the average higher than the DE estimates. This may be expected in the case of the LBDE and LBDiff which incorporate input from the difference estimator, which gives rise to estimates that are themselves higher than the DE estimates. The ratio estimator has the smallest R-MD and R-RMSD among the three list-based estimators. At the regional level, the performance of these estimators is fairly inconsistent for all of the labor items.
- The predicted NOL estimates compare favorably to the DE for all three items in the case of hired workers, as do the post-stratified estimates in the case of weekly hours and wage rates for hired workers. However, these two estimators are completely unreliable for the other two types of workers. Overall, these two estimators have R-RMSD substantially higher than other estimators.

VARIANCE ESTIMATION FOR DIRECT EXPANSION

The current variance estimate for the DE is computed in terms of expanded sample values for an item of interest and can be formulated as follows:

$$v_A = \sum_{h=1}^{H} s_h^2 \tag{15}$$

where

$$s_h^2 = \frac{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}{n_h - 1}, \tag{16}$$

$\hat{y}_i$ is the expanded value for the ith sampled PSU, and

$$\bar{\hat{y}} = \frac{\sum_i \hat{y}_i}{n_h}.$$

This variance computation is based on a large sample estimator and thus, it is an approximation to the actual variance. The estimates tend to be biased upward and slightly overestimate the variance of the DE. However, the estimates are fairly stable.

Kott (1990) derived a design unbiased variance estimator for the DE taking into account the two phase sampling design that NASS utilizes for its agricultural labor surveys. This estimator of the variance can be described as follows:

Table 2: **R-MD and R-RMSD for Alternative Estimators Relative to DE (US Level)**

Item Estimated	Estimator						
	Diff	LBDE	LBDiff	PNOL	PSAF	Ratio	ReWt
	R-MD(%)						
	Hired						
Total	2.8	0.9	3.4	3.4	-0.3	2.0	0.0
Weekly Hours	0.0	-0.6	-0.2	1.1	2.6	-0.3	-0.0
Wage Rate	-0.1	1.2	0.6	-1.1	0.2	0.3	0.0
	Self-Employed						
Total	7.3	4.2	6.9	11.5	-6.9	2.6	0.0
Weekly Hours	0.7	-5.4	-4.5	18.4	19.5	-0.8	-0.0
	Unpaid						
Total	7.1	0.9	3.4	29.1	10.9	5.8	0.0
Weekly Hours	0.1	-5.7	-4.7	3.2	5.2	-0.4	-0.0
	R-RMSD(%)						
	Hired						
Total	3.6	4.9	5.5	5.4	7.9	4.3	0.2
Weekly Hours	0.3	1.5	0.9	1.4	3.0	0.9	0.1
Wage Rate	0.3	1.9	1.2	1.3	1.8	1.0	0.0
	Self-Employed						
Total	8.3	5.6	7.8	13.2	10.0	3.6	0.7
Weekly Hours	0.9	8.6	7.8	19.6	20.5	2.7	0.5
	Unpaid						
Total	8.2	7.8	8.5	30.7	20.2	7.9	0.6
Weekly Hours	0.6	9.2	8.0	4.2	5.7	2.1	0.3

$$v_U = v_A + v_B + v_C \qquad (17)$$

where v_A is the approximate variance described above, v_B is the variance component for the restratified second phase sampling and v_C is the adjustment for unbiasedness. For further details, refer to Kott (1990).

The ratio of the two variance estimates given in Equations (15) and (17),

$$\frac{v_A}{v_U}, \qquad (18)$$

is computed to compare the approximate variance (v_A) to the unbiased variance estimator (v_U). The results at the U.S. level are given in Figure 1 on page 11. The ratio of the average variance from Equations

(15) and (17) across all survey periods is 1.02 for the hired workers, -4.19 for the self-employed workers and 1.53 for the unpaid workers. These results lead to the following conclusions:

- The currently used variance estimator (v_A) is at most 2% higher than the unbiased variance estimator (v_U) in the case of hired workers and mostly less than 20% higher than the unbiased variance estimator in the case of unpaid workers. An exception occurs in 1995-96 where the variance ratios are showing an unusual pattern.
- The unbiased variance estimator (v_U) is completely unreliable for the case of self-employed workers. Its value is negative in half of the survey periods considered.

Also, computed is the coefficient of variation for the DE using the current variance estimator (v_A), given by

$$\hat{\mathrm{CV}}_{\mathrm{de}} = \frac{\sqrt{v_A}}{\hat{Y}_{\mathrm{de}}}.$$

The results, obtained at the U.S. level, are displayed in Figure 2 on page 12. It can be seen in this figure that the $\hat{\mathrm{CV}}_{\mathrm{de}}$ranges approximately from 3.5% to 5.5% for total number of hired workers, from 2% to 3% for number of self-employed workers and from 3.5% to 7% for number of unpaid workers.

JACKKNIFE EVALUATIONS OF BIAS AND VARIANCE

The bias and variance of each estimator was evaluated using a replicated delete-a-random-group jackknife procedure as described in Appendix C. The jackknife replicates were developed by taking into account the sample design, then an estimate of the entire population was computed for each replicate by developing the necessary replicate weights. The jackknife procedure was carried out as follows:

1. For each of the list and area frame samples, 15 groups (called pseudo-replicates) were randomly created. This was accomplished by partitioning the first phase sample units in such a way that the sample design characteristics were preserved in each of the 15 groups of sample units.
2. The 15 jackknife replicates were formed by deleting one of the random groups at a time from the initial sample.
3. The sampling weights were recalculated for each jackknife replicate based on the new sample size.
4. A set of 15 jackknife estimates for each item was made for each estimator by applying the estimation process to each of the 15 sets of jackknife replicated data.
5. The average and the variance of the 15 jackknife estimates of each item were computed.

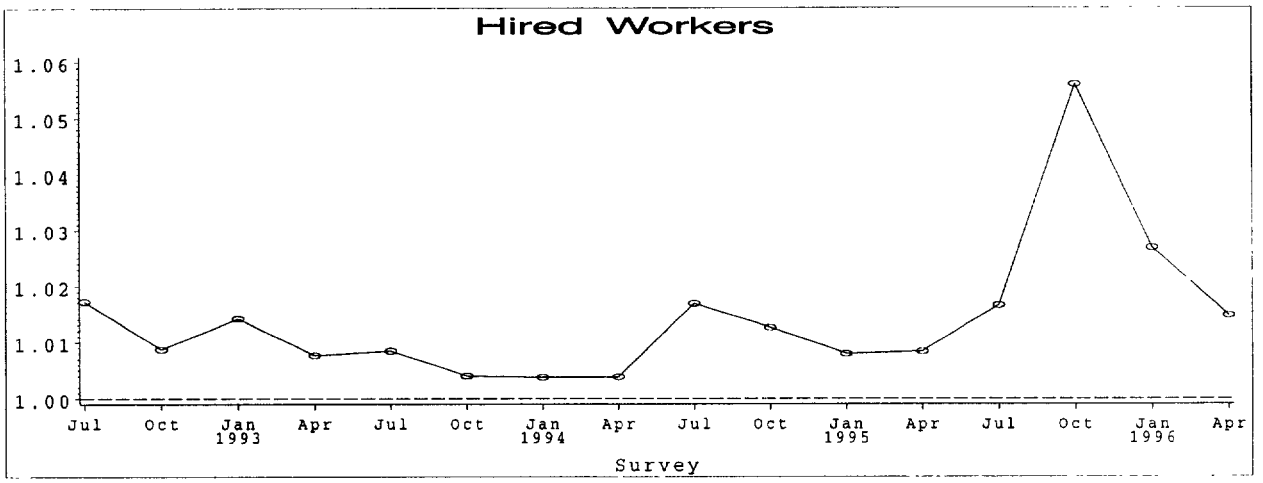
For an evaluation of bias, each estimator was compared to the direct expansion estimator in the following manner. An estimate of bias is obtained by

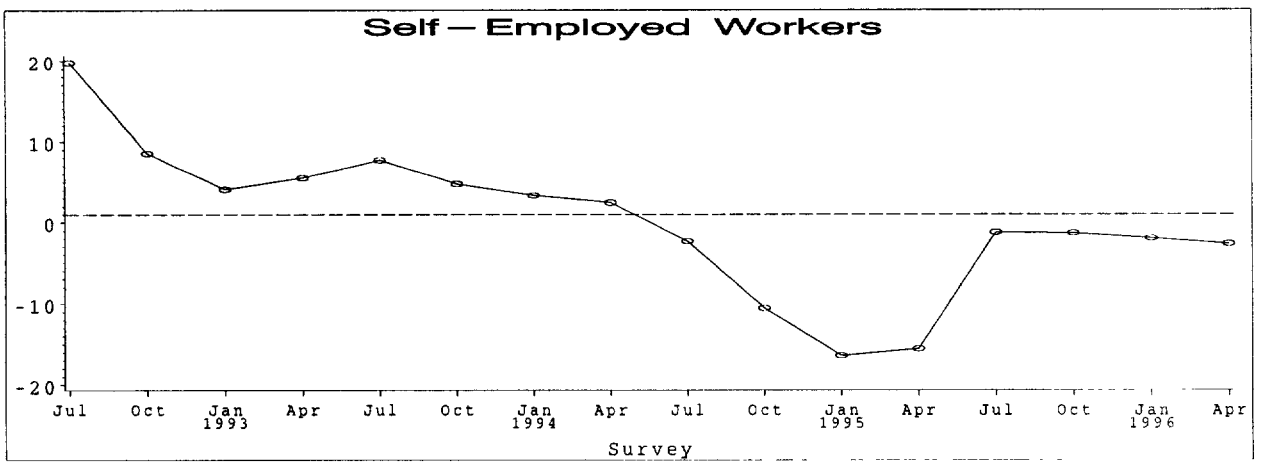
$$\hat{\mathrm{Bias}} = \hat{Y}_{\mathrm{est}} - \hat{Y}_{\mathrm{de}} \qquad (19)$$

where $\hat{Y}_{\mathrm{est}}$is the estimate using a particular estimator and $\hat{Y}_{\mathrm{de}}$is the corresponding estimate for the DE based on the complete sam-

Figure 1: Current Variance to Unbiased Variance Ratio for DE

United States Level





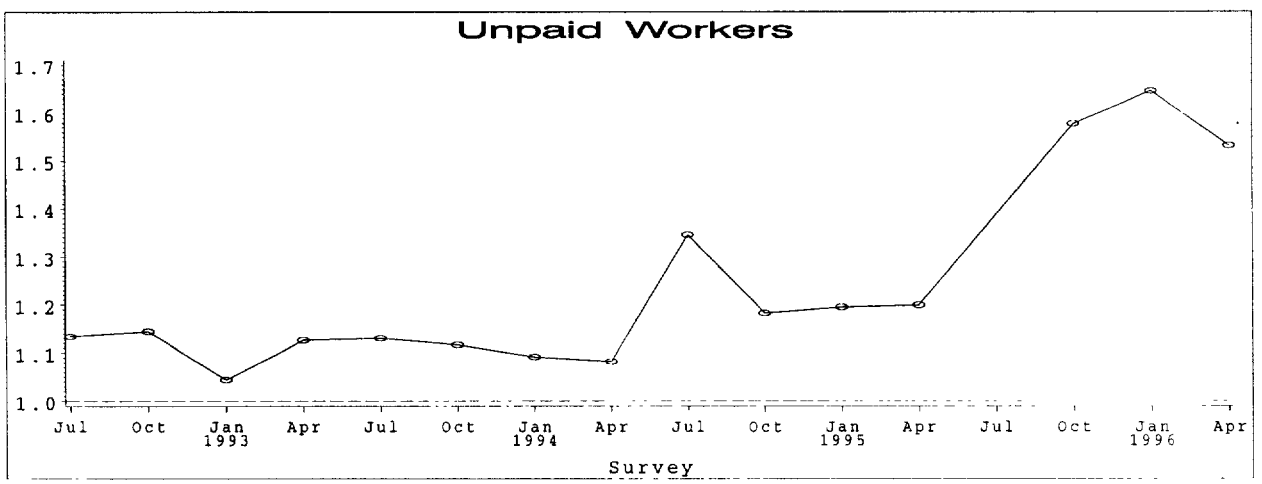


Figure 1:

Figure 2: Current Estimated CV for DE

United States Level

Figure 2:

ple data. Next, a 95% confidence interval for bias is computed by

$$\hat{\text{Bias}} \pm t_{14,0.025}\hat{\text{SD}} \tag{20}$$

where

$$\hat{\text{SD}} = \sqrt{\frac{14}{15}\sum_{i=1}^{15}(D_i - D)^2},$$

$$D_i = \hat{Y}_{J,i} - \hat{Y}_{\text{de},i}, \quad i = 1, 2, \ldots, 15,$$
$$D = \hat{Y}_{\text{est}} - \hat{Y}_{\text{de}},$$

and $\hat{Y}_{J,i}$ is the ith jackknife replicated estimate for the estimator being evaluated and $\hat{Y}_{\text{de},i}$ is the DE estimate corresponding to that jackknife replication

The bias estimates and confidence intervals, both computed at the U.S. level relative to the DE for the various estimators, are depicted in Figures 3-5 on pages 14-16. The following conclusions are drawn from these results.

- The alternative estimators do not exhibit any significant bias for number of hired workers.
- The predicted NOL estimator has positive bias in estimating number of unpaid workers (approximately 30% higher overall). The difference estimator shows bias of at most 10% higher than DE in estimating number of unpaid workers.
- All of the estimators except the reweighted show a slight, consistent bias (though mostly insignificant) in estimating number of self-employed workers. The bias is positive except in the case of the post-stratified estimator which has a negative bias.

Jacknife Derived Statistics

The mean and variance are computed using the 15 jackknife replicated estimates for each of the estimators evaluated here. Thus, the jackknife mean value is

$$\hat{\bar{Y}}_J = \frac{1}{15}\sum_{i=1}^{15}\hat{Y}_{J,i} \tag{21}$$

where $\hat{Y}_{J,i}$ is the labor item estimate based on the ith jackknife replicate. The jackknife bias (Quenoulli's estimate) is then given by

$$14(\hat{\bar{Y}}_J - \hat{Y})$$

where $\hat{Y}$ is the estimate for a labor item computed based on the complete sample data.

The usual jackknife estimate of the variance is obtained by considering the deviations of the jackknife replicated estimates from their mean. This estimate is given by

$$\hat{\text{V}}_{1,J} = \frac{14}{15}\sum_{i=1}^{15}(\hat{Y}_{J,i} - \hat{\bar{Y}}_J)^2. \tag{22}$$

A slightly more conservative estimate of the variance is obtained by considering deviations of the jackknife replicated estimates from the estimate based on the complete sample data. This estimate is given by

$$\hat{\text{V}}_{2,J} = \frac{14}{15}\sum_{i=1}^{15}(\hat{Y}_{J,i} - \hat{Y})^2. \tag{23}$$

As the number of jackknife replicates become large, the two variance estimates, which are referred to as "Variance 1" and "Variance 2", converge to the same value (see page 60 in Appendix C).

Figure 3: Relative Bias from DE with 95% C.I.

Hired Workers

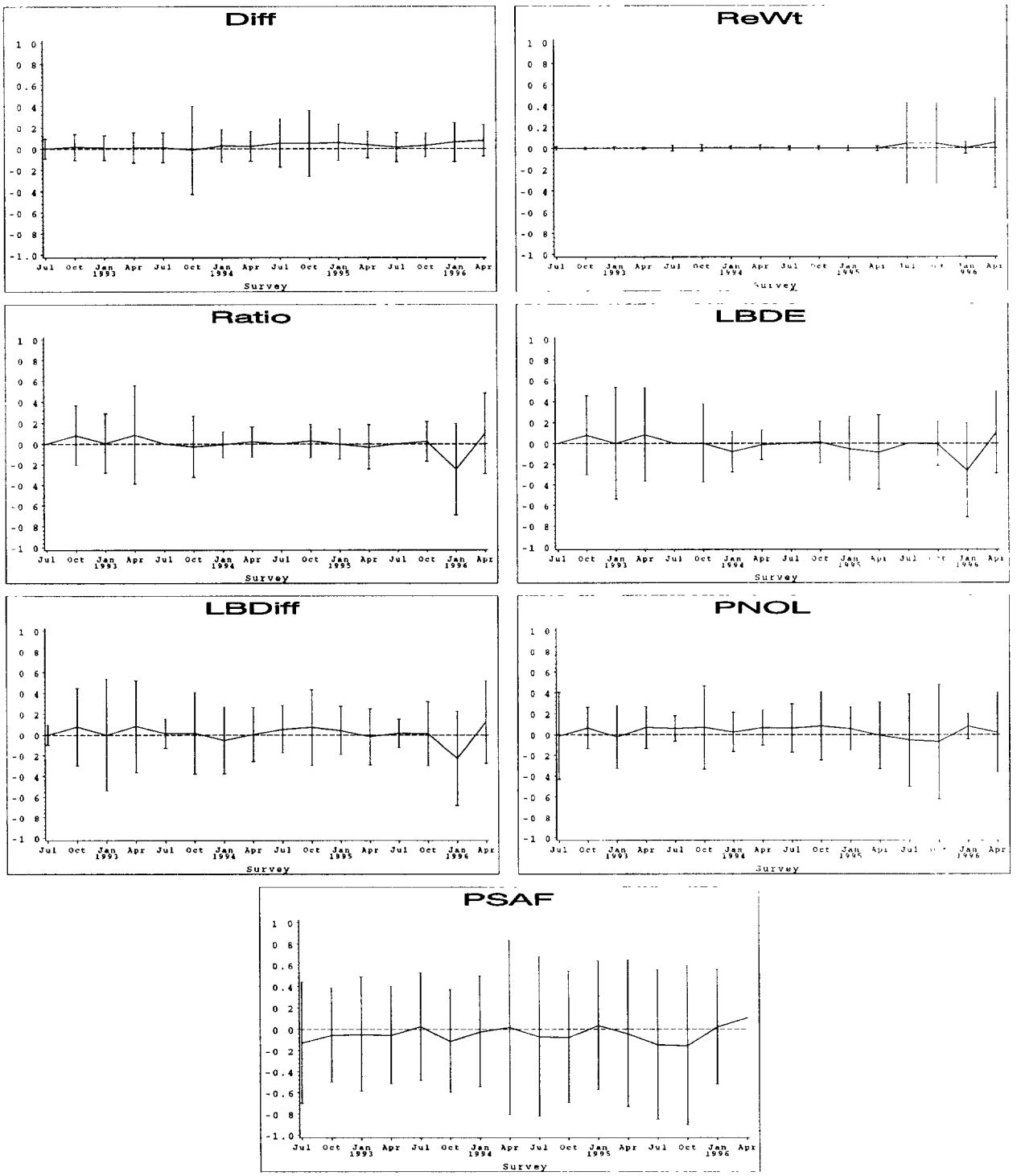


Figure 4: Relative Bias from DE with 95% C.I.

Self – Employed Workers

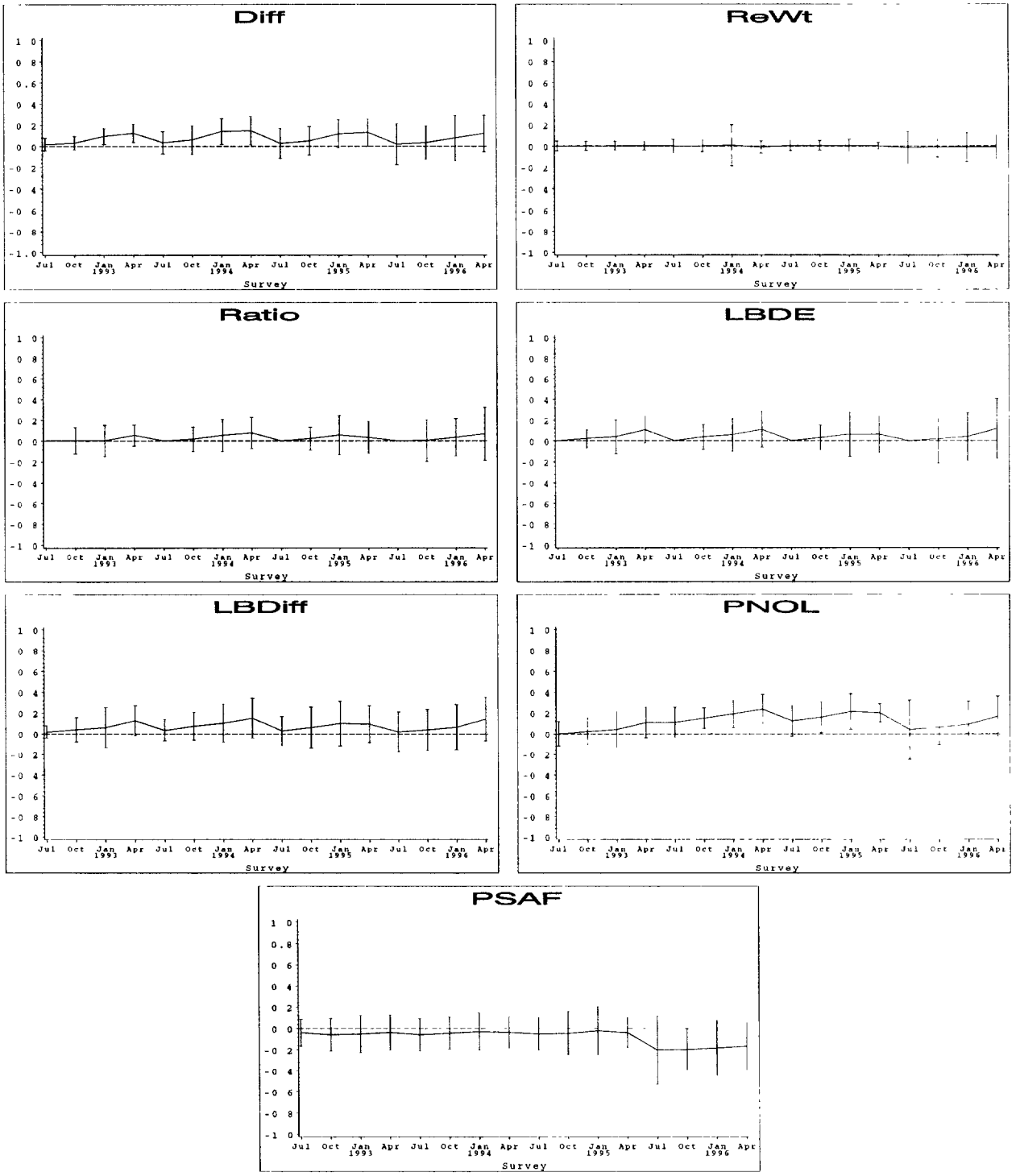
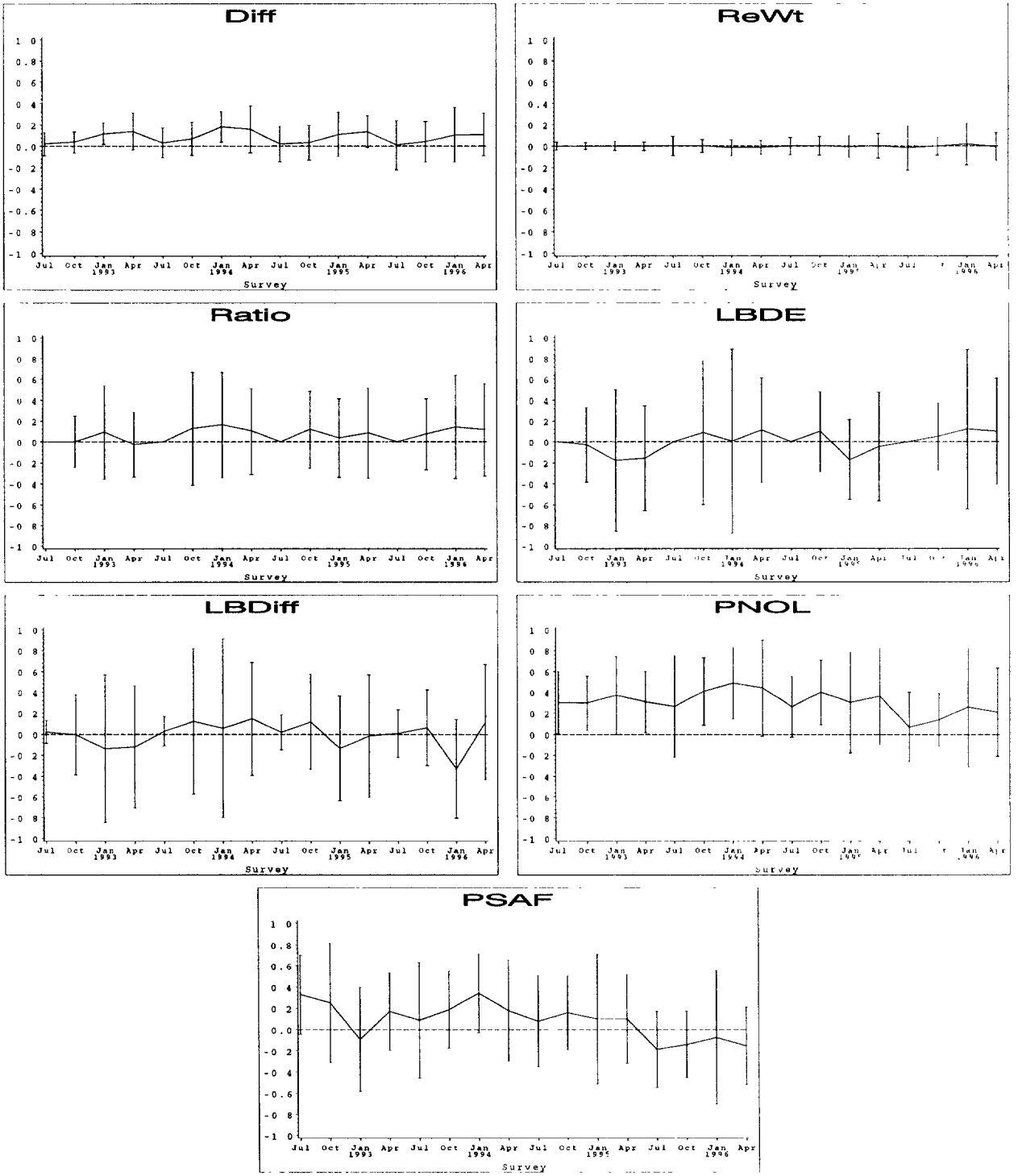


Figure 5: Relative Bias from DE with 95% C.I.

Unpaid Workers



As a measure of how conservative Variance 2 is with respect to Variance 1, the variance ratio,

$$\frac{\hat{V}_{2,J}}{\hat{V}_{1,J}}, \tag{24}$$

is computed. Table 3 on page 22 shows the average value of the variance ratio for each estimator over the 16 survey periods for number of hired, self-employed and unpaid workers.

From the table we see that the jackknife bias is generally small for all the estimators considered, since the variances ratio is normally 1.00. Thus, for consistency, Variance 2 is used in all the computations that follow.

The estimated coefficient of variation for an estimator is computed by

$$\hat{CV}_{2,J} = \frac{\sqrt{\hat{V}_{2,J}}}{\hat{\bar{Y}}_J}. \tag{25}$$

The following conclusions are drawn using Table 4 on page 23, which lists the average $\hat{CV}_{2,J}$ over the 16 survey periods for each estimator and each item estimated, and using Figure 6 on page 18, which displays the changes in $\hat{CV}_{2,J}$ over time for number of hired, self-employed and unpaid workers.

- $\hat{CV}_{2,J}$ for the DE is mostly the smallest, though it varies for farm labor item and survey period. The largest value of $\hat{CV}_{2,J}$ can be as much as two times the smallest $\hat{CV}_{2,J}$ value for an item estimated across survey periods.
- The $\hat{CV}_{2,J}$ values for the reweighted, ratio, difference and predicted NOL estimators are fairly consistent with the DE in the case of estimating total number of hired, self-employed and unpaid workers.
- The LBDE, LBDiff and post-stratified estimators display no specific pattern in their $\hat{CV}_{2,J}$ values. Consistency in $\hat{CV}_{2,J}$ values may exist over time for some items.

A 95% confidence interval for the relative jackknife bias is

$$\frac{(\hat{\bar{Y}}_J - \hat{Y}) \pm t_{14,0.025}\sqrt{\hat{V}_{2,J}}}{\hat{Y}}. \tag{26}$$

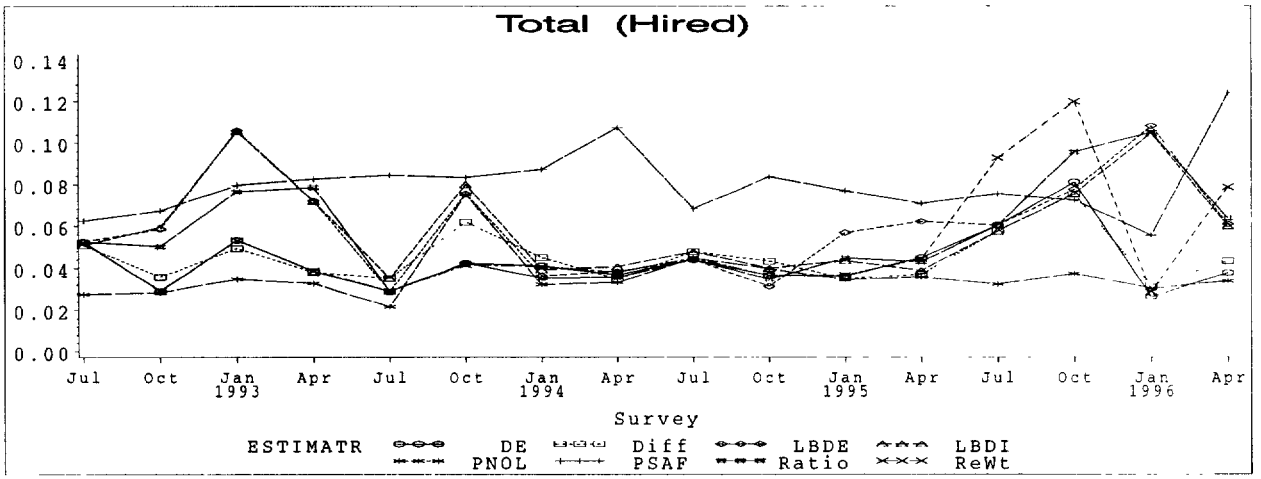
The confidence intervals are plotted for each estimator for total number of hired, self-employed and unpaid workers in Figures 7-9 on pages 19-21.

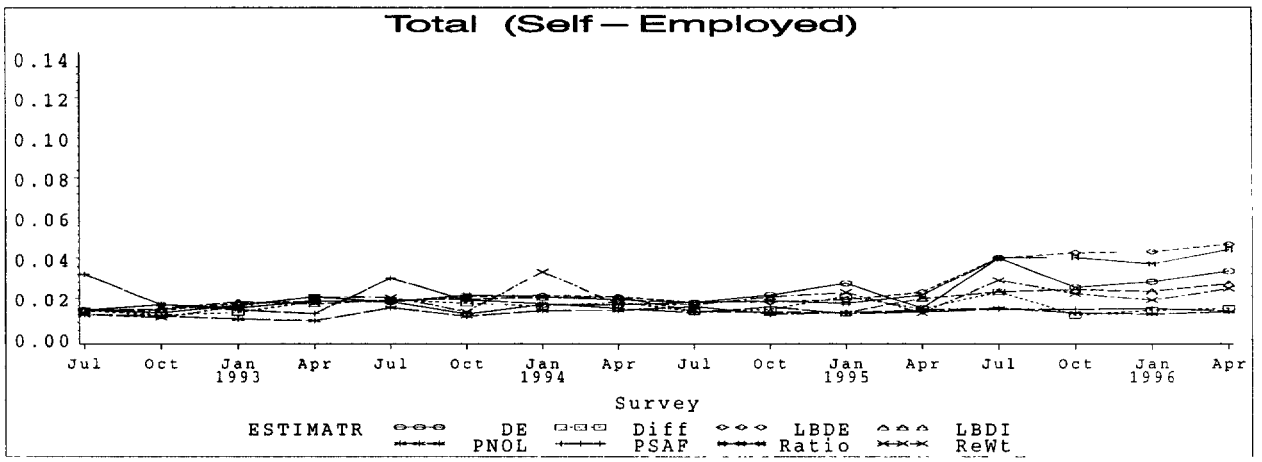
In terms of the jackknife bias evaluations, none of the estimators except the post-stratified estimator show any bias. The post-stratified estimates are indicative of negative bias during 1992-94, but for later years the post-stratified estimator behaved as expected. The reason for this is not clearly understood.

However, it should be noted that the list frame design strata were defined differently in 1992-94 than they were in 1994-96 when they were essentially the same as the poststrata. The design strata in 1992-94 were quite different than the poststrata. It is quite possible that these design changes caused the differences in the jackknife biases between 1992-94 and 1994-96. It should also be noted that the jackknife replicate weights used in deriving the biases for the post-stratified estimates were computed with a different program than those with the other estimators. Even though the programs were examined independently at NASS and Houston and no errors were found, it is still possible that some subtle differences in the pro-

Figure 6: CV using Variance 2

United States Level





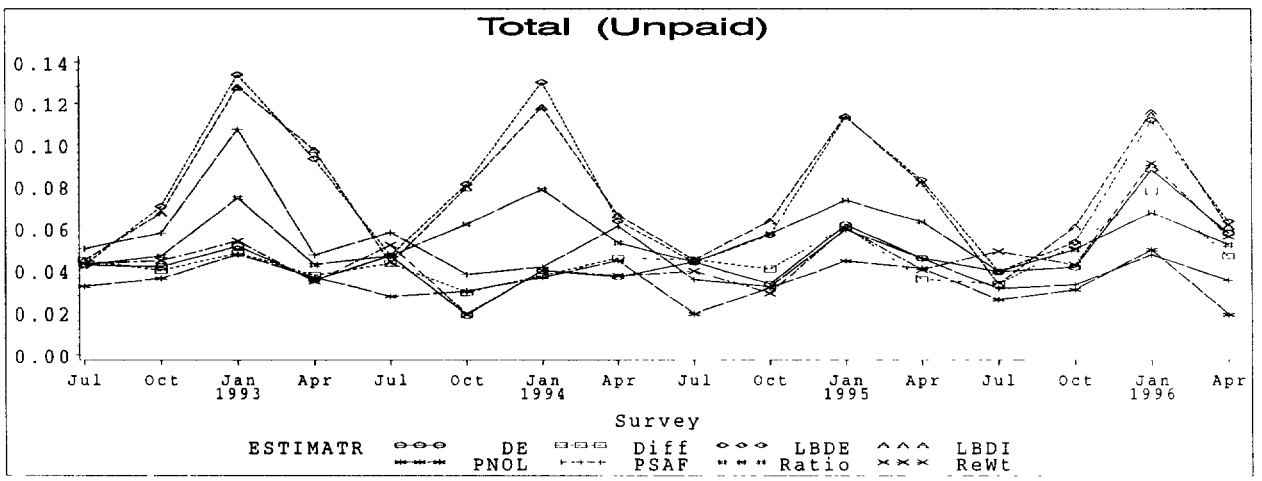


Figure 7: Jackknife Relative Bias with 95% C.I.

Hired Workers

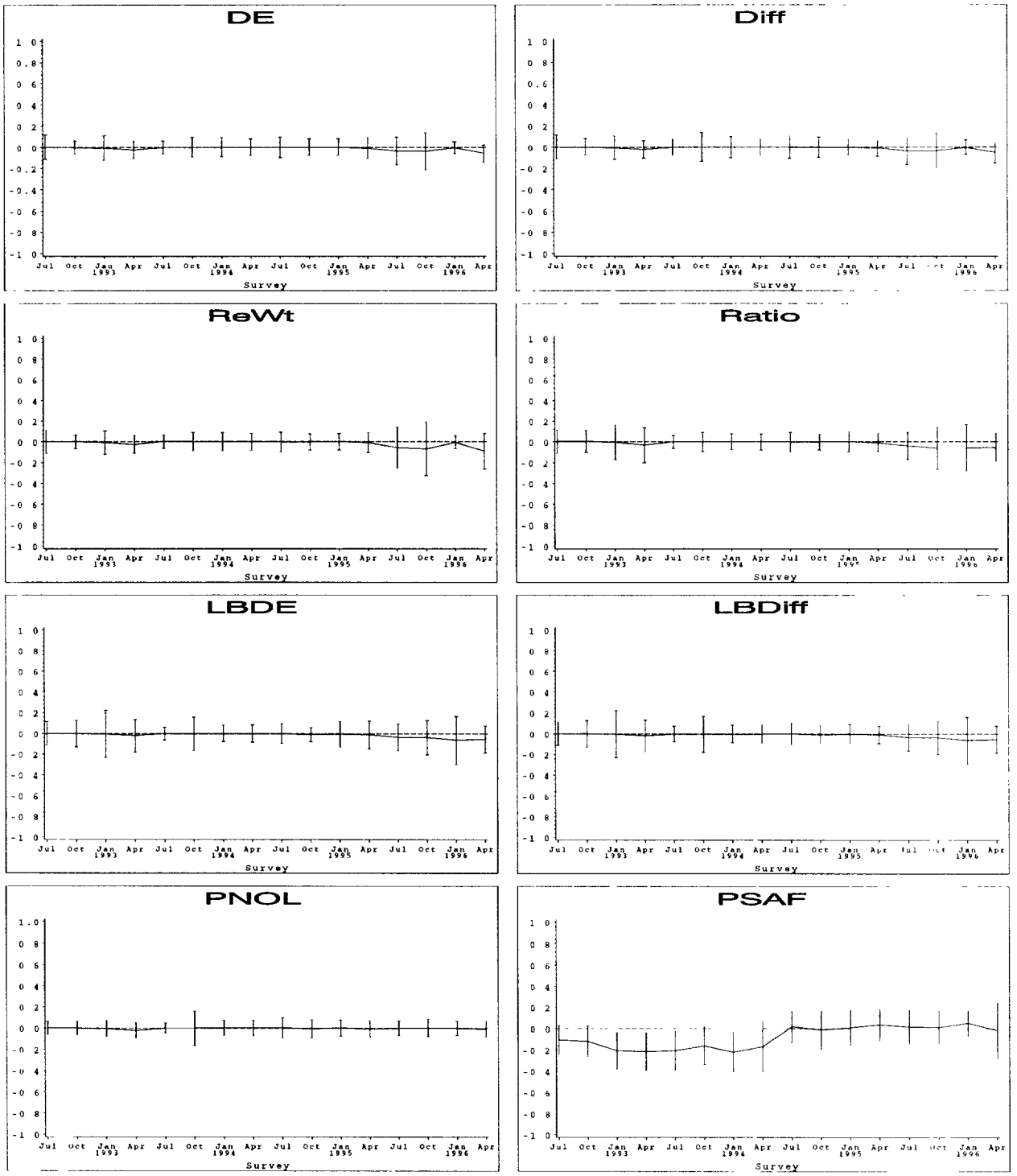


Figure 8: Jackknife Relative Bias with 95% C.I.

Self – Employed Workers

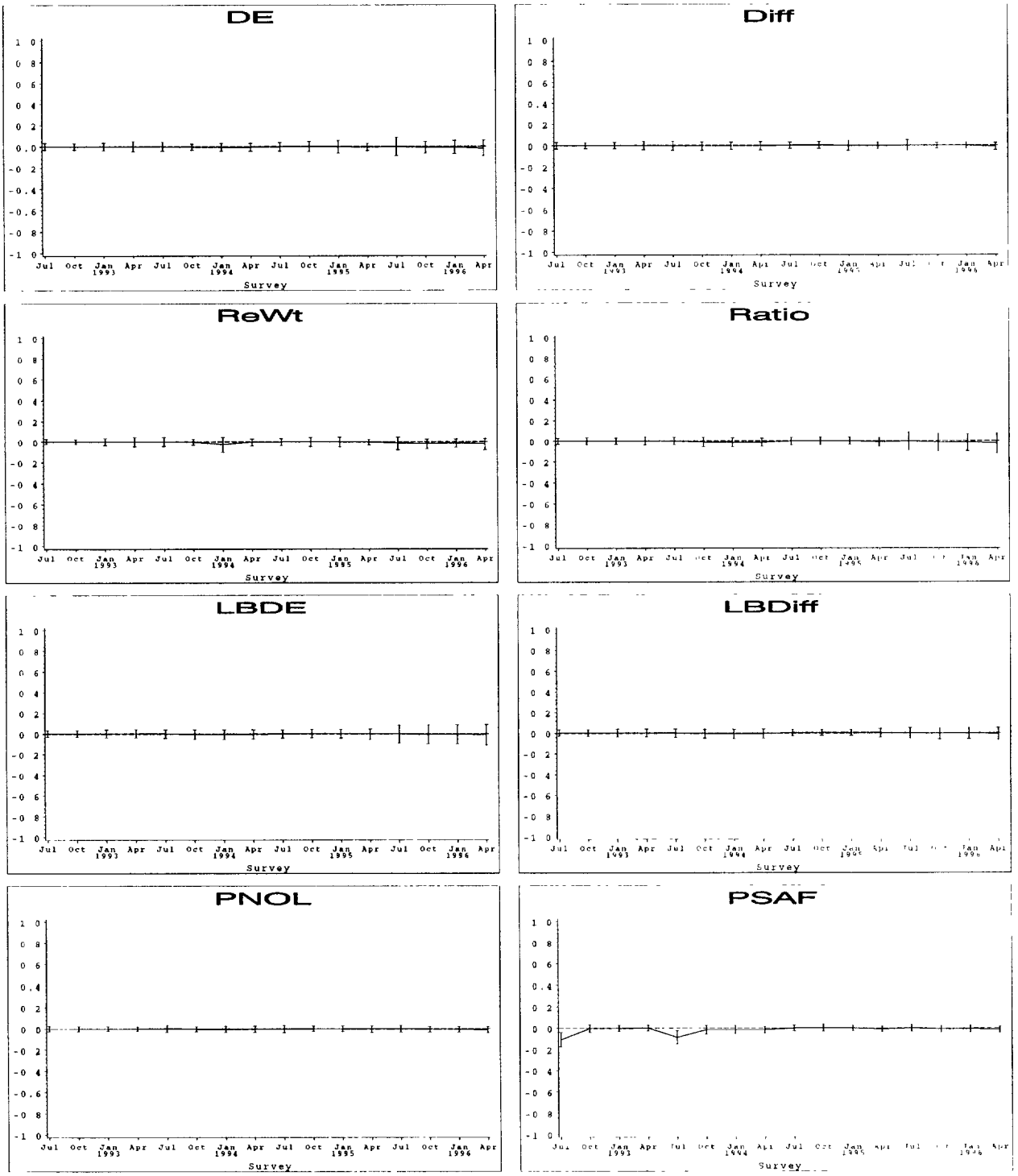


Figure 9: Jackknife Relative Bias with 95% C.I.

Unpaid Workers

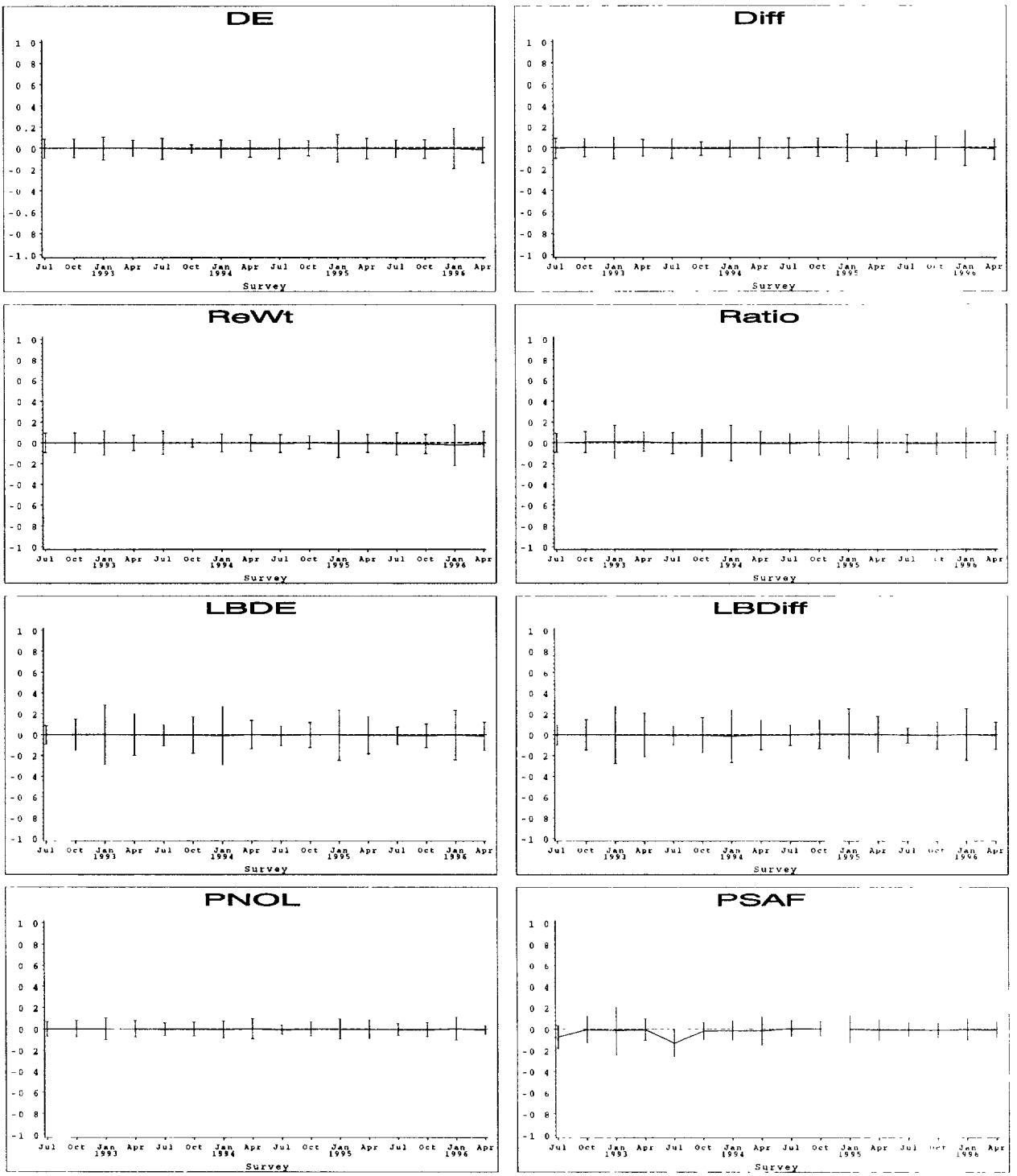


Table 3: **Average Ratio of Variance 2 to Variance 1 at US Level**
(Standard deviation of average ratio is in parenthesis)

Item Estimated	DE	Diff	LBDE	LBDIFF	PNOL	PSAF	RATIO	ReWt
Hired:								
Total	1.00	1.00	1.00	1.00	1.00	1.30	1.00	1.00
	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(0.37)	(0.00)	(0.01)
Weekly Hours	1.00	1.00	1.00	1.00	1.00	1.60	1.00	1.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(1.87)	(0.01)	(0.01)
Wage Rate	1.00	1.00	1.00	1.00	1.00	1.33	1.00	1.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(1.14)	(0.00)	(0.00)
Self-Employed:								
Total	1.01	1.00	1.00	1.00	1.01	1.40	1.01	1.01
	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(1.14)	(0.01)	(0.02)
Weekly Hours	1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)
Unpaid:								
Total	1.00	1.00	1.00	1.00	1.00	1.05	1.00	1.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.15)	(0.00)	(0.00)
Weekly Hours	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)

grams caused the differences observed in the 1992-94 and 1994-96 bias.

In order to evaluate the variance of an estimator relative to that of the DE, we computed the variance ratio given by:

$$\frac{\hat{V}_{2,J}}{V_{2,\mathrm{DE}}}. \tag{27}$$

The results are displayed in Table 5 on page 24 in the form of the average ratio over the 16 survey periods for each estimator and each item estimated.

Some of the alternative estimators have a substantially larger variance than that of the DE. The variance ratio results in Table 5 on page 24 show that, compared to the DE, the reweighted expansion has variance as much as 21% higher than the DE, the difference estimator has variance as much as 29% higher than the DE, and the predicted NOL has variance as much as 41% higher than the DE. For each of the other four estimators (the LBDE, LBDiff, ratio and the post-stratified area frame) the variance is as much as 2 to 4 times higher than the DE. However, in most cases, the reweighted, difference and predicted NOL estimators have

Table 4: **Average $\hat{CV}_{2,J}$ at US Level**
(Standard deviation of average $\hat{CV}_{2,J}$ is in parenthesis)

Item Estimated	DE	Diff	LBDE	LBDiff	PNOL	PSAF	RATIO	ReWt
Hired:								
Total	0.04	0.04	0.05	0.05	0.04	0.08	0.05	0.04
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)
Weekly Hours	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Wage Rate	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Self-Employed:								
Total	0.02	0.02	0.03	0.02	0.01	0.02	0.02	0.02
	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
Weekly Hours	0.02	0.02	0.03	0.03	0.01	0.01	0.02	0.02
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)
Unpaid:								
Total	0.05	0.05	0.08	0.08	0.03	0.05	0.06	0.05
	(0.00)	(0.01)	(0.03)	(0.03)	(0.01)	(0.02)	(0.01)	(0.02)
Weekly Hours	0.03	0.03	0.06	0.05	0.02	0.02	0.04	0.03
	(0.00)	(0.01)	(0.03)	(0.03)	(0.00)	(0.00)	(0.02)	(0.01)

variance ratios close to 1; hence, they can be regarded as being as efficient as the DE.

CONCLUSIONS

A number of conclusions follow as a result of extensive evaluation of the potential approaches that were investigated in the development of efficient list-based or list-only estimators for the Quarterly Agricultural Labor Surveys.

First, the farm-type variable was seen to be useful in post-stratification of samples from both the labor list and the JAS. Second, a number of data anomalies were observed with respect to farm value of sales and number of self-employed and unpaid workers. The farm value of sales was thought to be a reasonable predictor for the farm labor; however, the labor survey data did not show this to be true.

The current direct expansion estimates tend to be higher than the official statistics. Overall, the DE estimates are approximately 4 percent higher than the Board.

Since this study was comprised of three

Table 5: **Average Ratio of Variance 2 to DE Variance at US Level**
(Standard deviation of average ratio is in parenthesis)

Item Estimated	DE	Diff	LBDE	LBDIFF	PNOL	PSAF	RATIO	ReWt
Hired:								
Total	1	1.26	2.46	2.62	1.02	4.82	1.95	1.01
	(0)	(0.39)	(2.48)	(2.70)	(0.78)	(4.12)	(1.96)	(0.03)
Weekly Hours	1	1.05	1.98	1.57	0.72	3.54	1.32	1.01
	(0)	(0.18)	(1.58)	(1.28)	(0.34)	(7.57)	(0.91)	(0.07)
Wage Rate	1	1.04	1.99	1.39	0.95	3.71	1.40	1.00
	(0)	(0.22)	(1.96)	(1.02)	(0.52)	(3.37)	(0.97)	(0.02)
Self-Employed:								
Total	1	0.88	1.66	1.20	0.63	1.02	1.50	1.21
	(0)	(0.56)	(0.87)	(0.75)	(0.39)	(1.25)	(0.78)	(1.26)
Weekly Hours	1	1.06	2.20	2.67	0.76	1.35	1.51	1.10
	(0)	(0.19)	(1.08)	(1.39)	(0.48)	(1.18)	(0.83)	(0.80)
Unpaid:								
Total	1	1.29	4.22	4.40	1.41	2.23	2.62	1.07
	(0)	(0.55)	(5.42)	(5.46)	(1.24)	(2.25)	(3.27)	(0.14)
Weekly Hours	1	0.97	5.20	5.00	1.07	1.36	2.32	0.98
	(0)	(0.15)	(5.21)	(4.78)	(0.88)	(1.19)	(2.24)	(0.16)

categories of estimators, the conclusions about the estimators are described separately for each category.

Quarterly Labor NOL Samples Used

These estimators, which include the currently employed direct expansion estimator, the reweighted expansion estimator and the difference estimator, require the use of quarterly NOL labor samples. Overall, the reweighted expansion estimator compares well with the DE, but exhibits more variability as expected. For example, in the case of self-employed workers, the reweighted estimator has variance as much as 21 percent higher than the DE. Both estimators compare equally well with the Board estimates at the U.S. level. The difference estimator uses information from the JAS in an attempt to better predict the current NOL. It is more difficult to implement, computationally intensive and requires the same amount of data as the other two. Moreover, the difference estimator does not perform as well as the DE, thus has no advantage over the DE.

July Labor NOL Samples Used

The list-based direct expansion, list-based difference and ratio estimators only require NOL samples in the July quarterly labor survey. No samples are required from the NOL

in the other quarters. The ratio estimator, which is obtained by multiplying the current list estimate by the ratio of the July total to July list estimates, is the most consistent of the three list-based estimators. The list-based direct expansion and difference estimators exhibit more variability and slightly more bias than the ratio estimator in some cases.

No NOL Samples Used

The predicted NOL and post-stratified area frame estimators do not require samples from the NOL in any quarterly labor survey. Although, the predicted NOL performs considerably better than the post-stratified estimator, it does not perform at an acceptable level. However, it has the best performance among the list-based and list-only estimators for the hired and unpaid workers, but it performs very poorly for the self-employed workers. The post-stratified area frame estimator performs very poorly in the cases of total self-employed and unpaid workers as well as average weekly hours for each type of farm labor.

Overall, none of the estimators consistently matched the performance of the DE as shown by the jackknife evaluations. Although the ratio estimator had $\hat{CV}_{2,J}$ estimates very similar to those of the DE, it was slightly biased compared to the DE in some quarterly survey periods.

FURTHER REMARKS

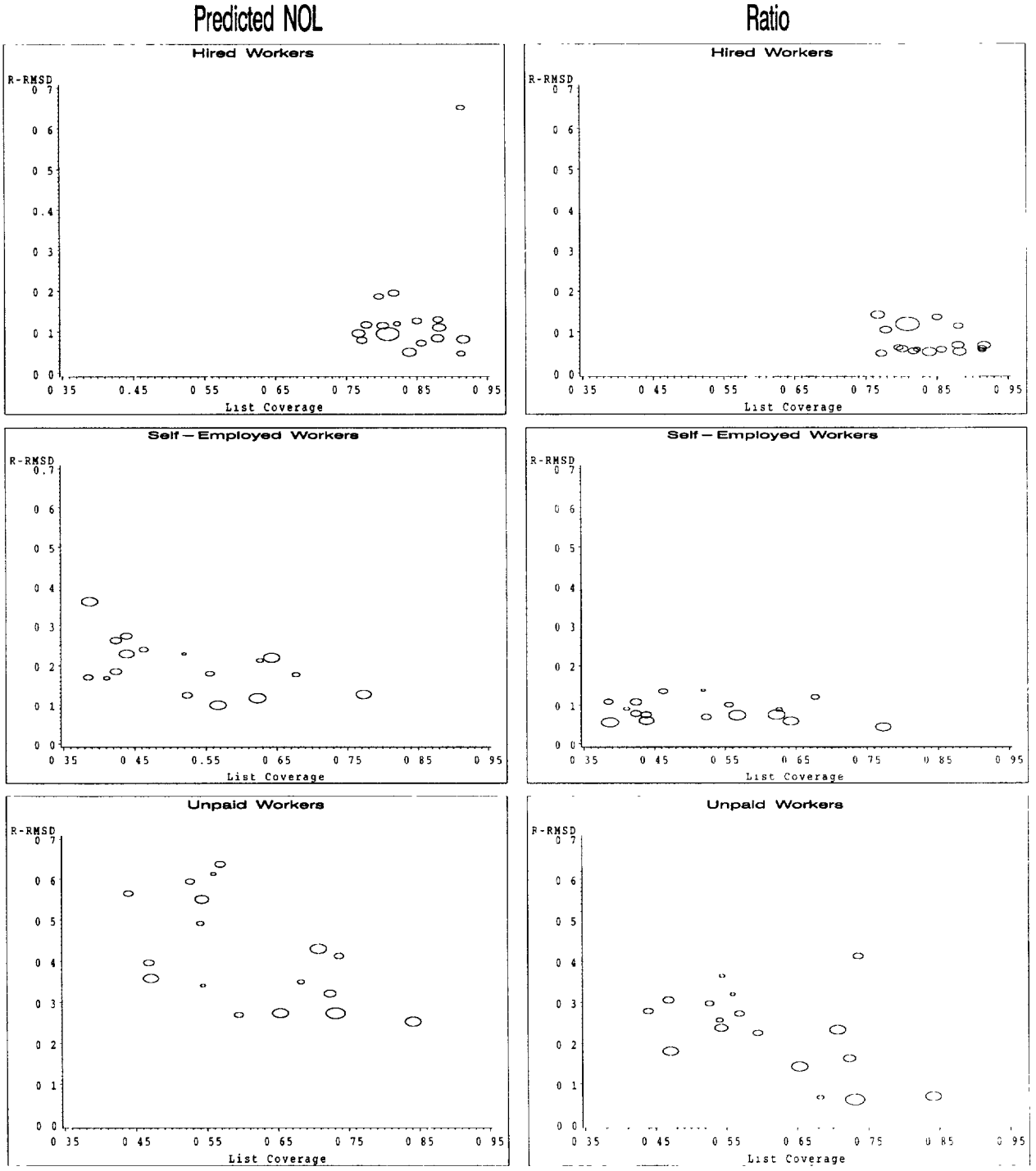
As indicated earlier, this study shows that as an alternative to the DE, one may consider using the ratio estimator if only July labor NOL samples are used, or the predicted NOL estimator if no NOL samples are to be used in estimation. The performance of these two estimators is examined as a function of the list frame coverage. In Figure 10 on page 26, R-RMSD is plotted against the ratio of DE list estimate to DE total estimate for the number of workers (hired, self-employed and unpaid) for the 17 regions. Each region is represented by a circle where the size of the circle is proportional to the contribution of that region to the total DE estimate. The following conclusions can be drawn from the graphs:

1. The list frame coverage estimated by the ratio of the DE list estimate to the DE total estimate is substantially higher (can be as much as two times) for the hired workers than the self-employed or the unpaid workers.

2. R-RMSD is substantially higher for the self-employed and unpaid workers than for the hired workers, except when the list frame coverage exceeds 75 percent, where the performance of each of these two estimators is about the same for the three labor items estimated.

The number of partners was used as a regressor for the self-employed workers; however the two are not well correlated as can be seen from Table A2 and the scatter plots in Appendix A for self-employed workers. No regressor was used in the case of unpaid workers due to the lack of a reasonably correlated auxiliary variable. Thus, the poor performance of the predicted NOL estimator for self-employed and unpaid workers is apparently due to lack of a sufficiently correlated auxiliary variable similar to the peak number of workers in the case of hired workers.

Figure 10: Relative Root Mean Square Deviation from DE versus List Coverage



Note: The size of each circle is proportional to the contribution of the associated region to the total DE estimate

RECOMMENDATIONS

This study leads us to make the following recommendations.

1. The DE estimator should continue to be used until a list frame with sufficiently large (at least 75 to 85 percent) coverage for all farm labor items (hired, self-employed and unpaid) in each region is available.
2. The sample design should take into consideration not only peak number of workers, but also farm type and farm value of sales to improve sampling efficiency. One way this can be accomplished is to do the following: 1) stratify the list frame records as is currently done, which is by peak number of workers when available or by farm value of sales and a common and uncommon farm type classification when peak number of workers is not available, 2) sort the records within each stratum by farm type and farm value of sales, and then 3) select a systematic sample within each stratum. This procedure will produce a smooth evenly spread distribution of the sample by farm type and farm value of sales within each of the current strata. In effect, it will randomize the sample size proportionally to all farm type and farm value of sales categories. Thus, it can only improve current sampling efficiency.
3. A research effort should be made toward the development of statistically defensible procedures for handling highly influential, unexpected data records.
4. Research efforts should be directed toward identifying auxiliary information sufficiently correlated with each of the labor items such as the peak number of self-employed workers and the peak number of unpaid workers (in addition to the currently available peak number of hired workers) to use in sample design or as regressors to further increase the efficiency of survey estimates.

REFERENCES

Allen, Rich (1997). "Statistical Defensibility as Used by U.S. Department of Agriculture, National Agricultural Statistics Service." *Journal of Official Statistics*, **8:4**, 481-498.

Chhikara, Raj S., Perry, Charles R., Deng, Lih-Yuan, Iwig William C., Spears Floyd M. and Cowles, Susan (1995). "Post-Stratification and Efficient Estimation in U.S. Agricultural Labor Surveys". ASA Proceedings of Survey Research Methods, 1995.

Kott, P.S. (1990). "Some Mathematical Comments on Modified Strawman Estimators". NASS Internal Discussion Paper.

Kott, P.S. (1990). "Mathematical Formulae for the 1989 Survey Processing System (SPS) Summary". USDA-NASS SRB Research Report No. SRB-90-08, May, 1990.

Perry, Charles R., Chhikara, Raj S., Deng, Lih-Yuan, Iwig, William C., and Rumburg, Scot (1993). "Generalized Post-stratification Estimators in the Agricultural Labor Survey", SRB Research Report No. SRB-93-04, Washington, D.C., July, 1993.

Rumburg, Scot, Perry, Charles R., Chhikara, Raj S., and Iwig, William C. (1993).

Analysis of a Generalized Post-Stratification Approach for the Agricultural Labor Survey. SRB Research Report No. SRB-93-05, July, 1993.

Shao, Jun and Dongsheng, Tu (1995). *The Jackknife and Bootstrap.* Springer-Verlag, New York.

Spears, Floyd M., Chhikara, Raj S. , Perry, Charles R., and Cowles, Susan (1997). "An Evaluation of Alternative Usda Agricultural Labor Survey Estimators". Proceedings of Survey Research Methods, Joint Statistical Meetings, 1997.

Spears, Floyd M., Chhikara, Raj S., Perry, Charles R., Iwig, William C. and Cowles, Susan (1996). "Agricultural Labor Estimation Using Only List-Frame Sampling". Proceedings of Survey Research Methods, Joint Statistical Meetings, 1996.

Vogel, Frederic A. (1995) "The evolution and Development of Agricultural Statistics at the United States Department of Agriculture." *Journal of Official Statistics*, **11:2**, 161-180.

Vogel, Frederic A. (1990). "Strawman Proposal for Multiple Frame Sampling" NASS Internal Memos, October-November.

Wolter, K.M. (1995). *Introduction to Variance Estimation.* Springer-Verlag, New York.

APPENDIX A: SCATTER PLOTS AND CORRELATION TABLES

List of Figures

List of Tables

Modeling

Each of the response variables $y_1, y_2, \ldots, y_7$ was plotted against the potential auxiliary variables x_1, x_2, x_3, and x_4 at the post-strata level. Given here are a set of scatter plots showing "y vs x" as well as "y/x vs x" for the case of number of hired workers and the number of self-employed workers for a selected number of post-strata in the regions of Appalachian I (comprised of VA and NC), California, Florida and North Plains (comprised of KS,NE,ND and SD).

Non-intercept Model

When a scatter plot of the variable y values versus the auxiliary variable x values shows a fitted line passing through the origin, it is indicative of a non-intercept linear model. Next, if a scatter plot of the variable ratio y/x versus x shows a random pattern about the regression coefficient line, it suggests that the model error is proportional to x^2. For further discussion on this modeling aspect, see Sarndal, et al (1992).

The two scatter plots for each post-stratum for the various cases given here support the assertion of a non-intercept linear model given by

$$y = \beta x + \epsilon$$

where $\mathrm{E}[\epsilon|x] = 0$ and $\mathrm{Var}[\epsilon|x] = \sigma^2 x$.

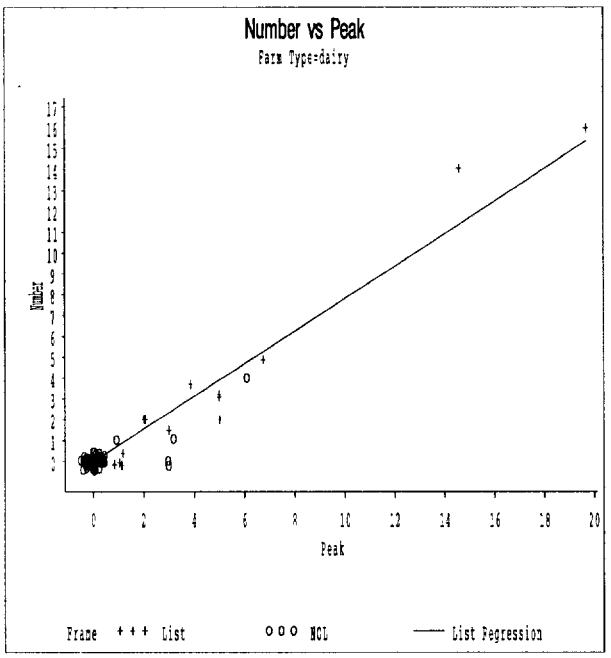
Applicability of Models to NOL

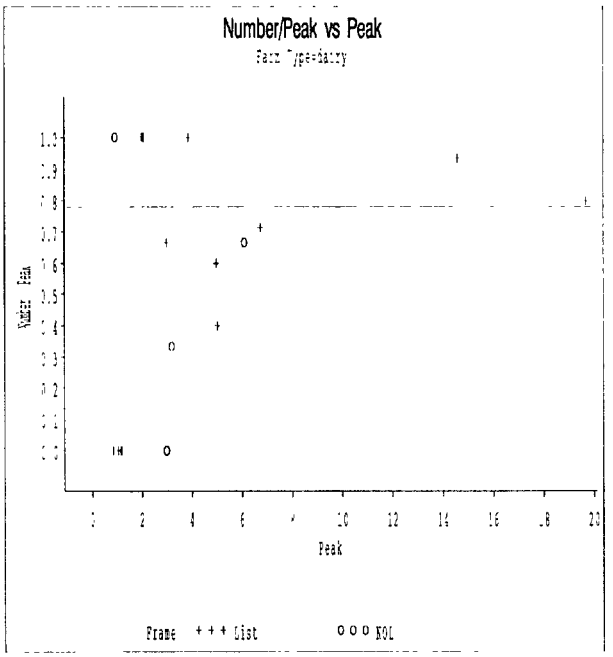
The correlation tables given here contrast the correlation coefficients obtained for the list frame samples and those for the NOL area samples across each post-strata for a selected number of cases. Apart from some exceptions, the correlations are comparable for the two frame samples. This outcome along with the scatter plots which include all list and NOL sample data suggests the applicability to NOL of the regression fits obtained using only list frame samples.

It is noted that the correlations for the self-employed workers, shown in Table A2, are not consistently good as compared to those in Table A1 for the hired workers. So, the number of partners is not a good predictor of the number of self-employed workers.

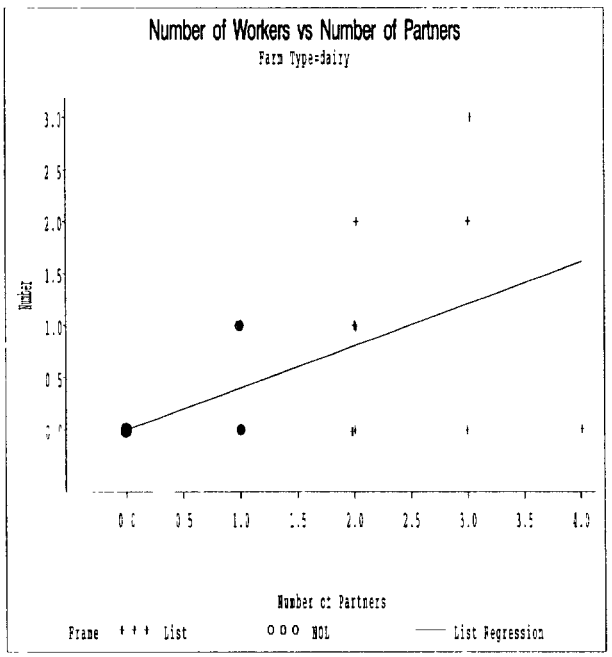
Figure A1: Appalachian I

Hired Workers





Self – Employed Workers



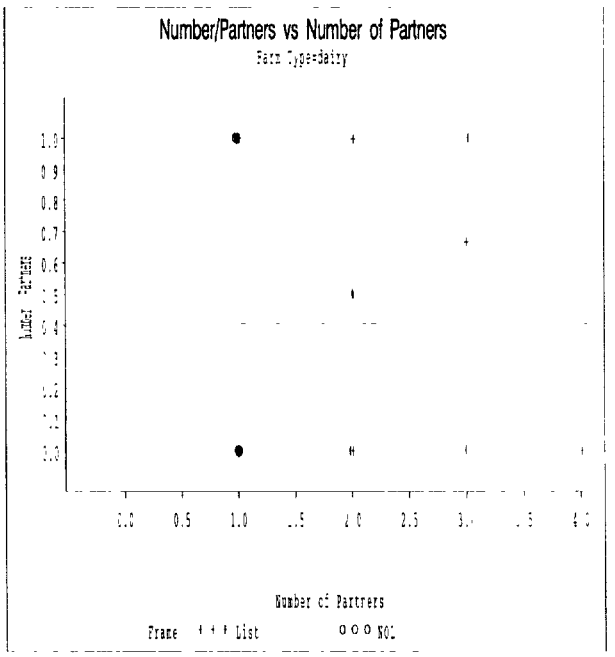
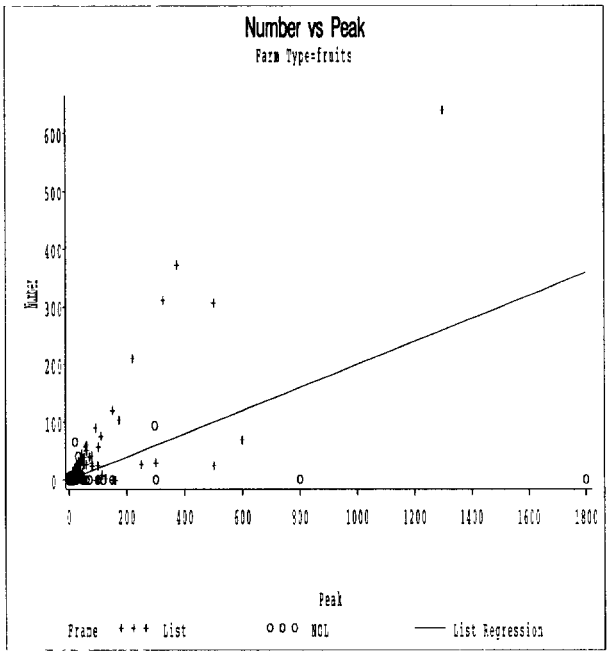
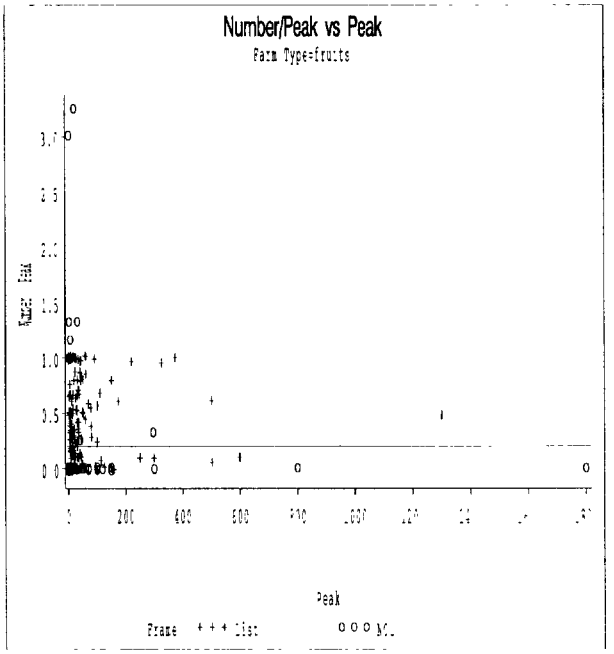


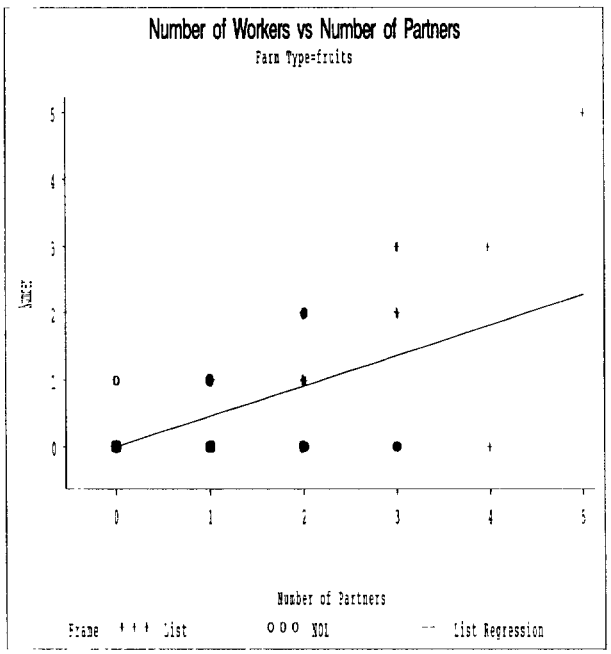
Figure A2: California

Hired Workers





Self – Employed Workers



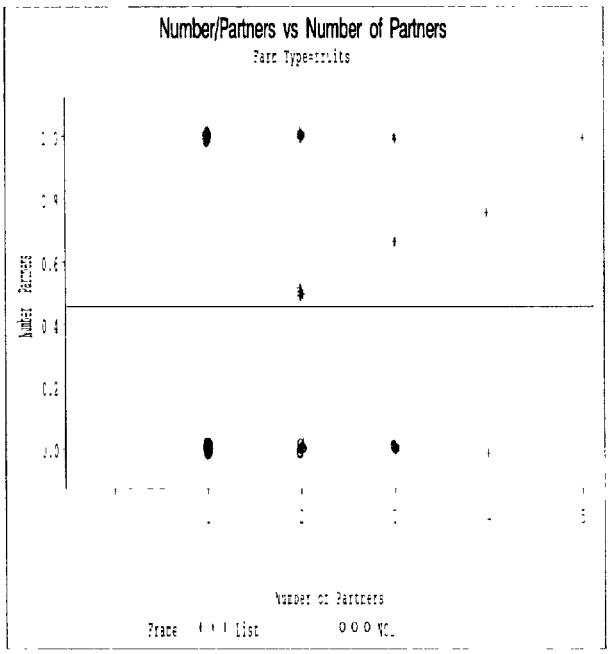
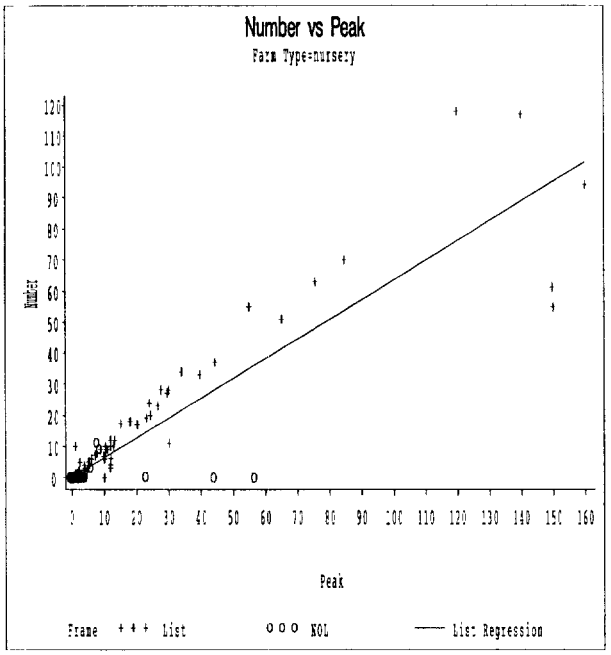
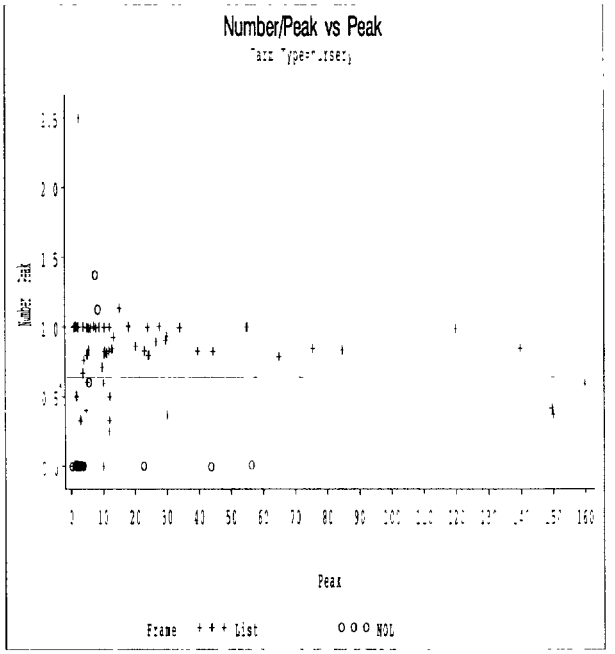


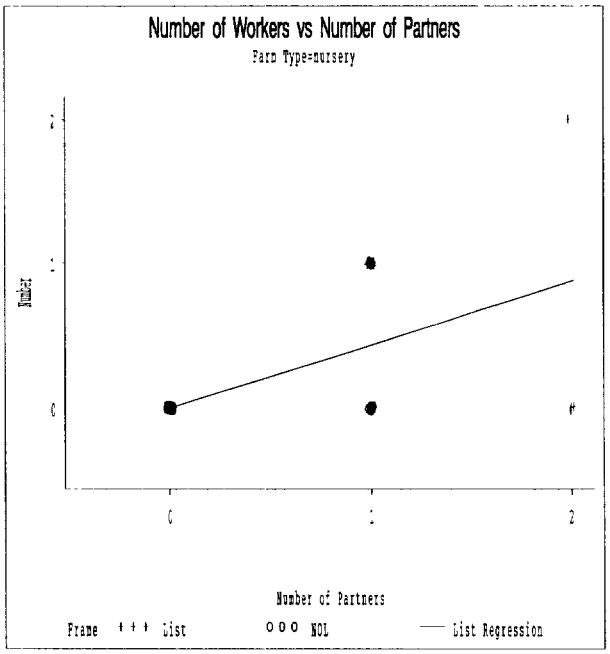
Figure A3: Florida

Hired Workers





Self – Employed Workers



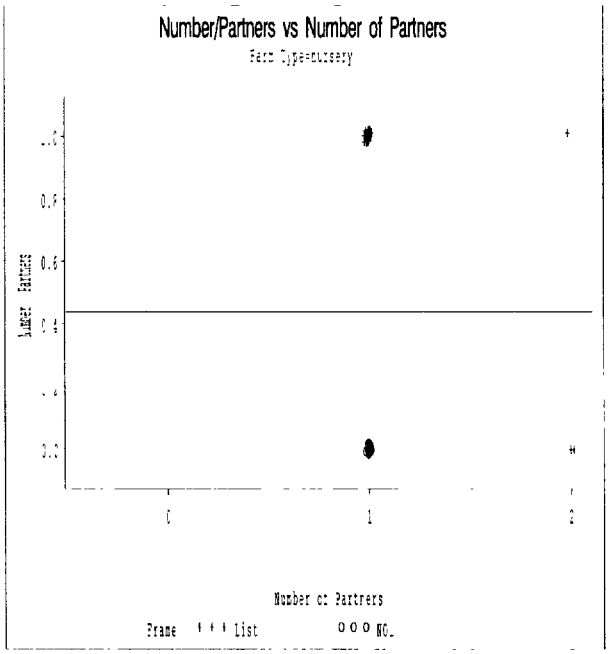
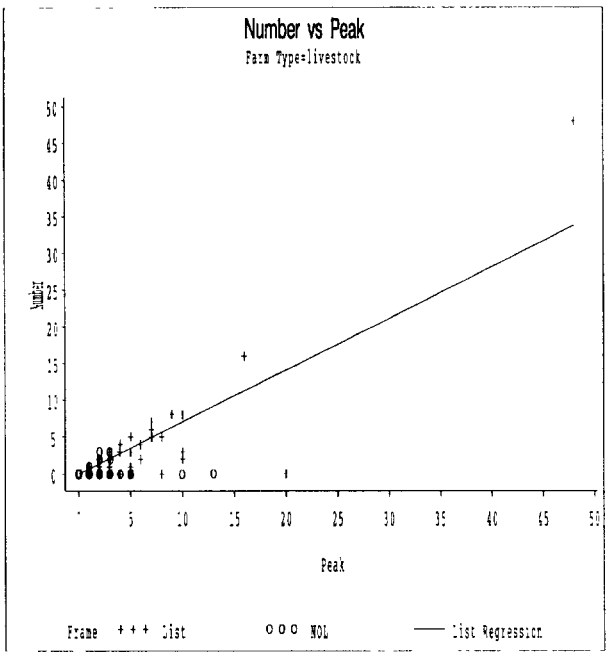
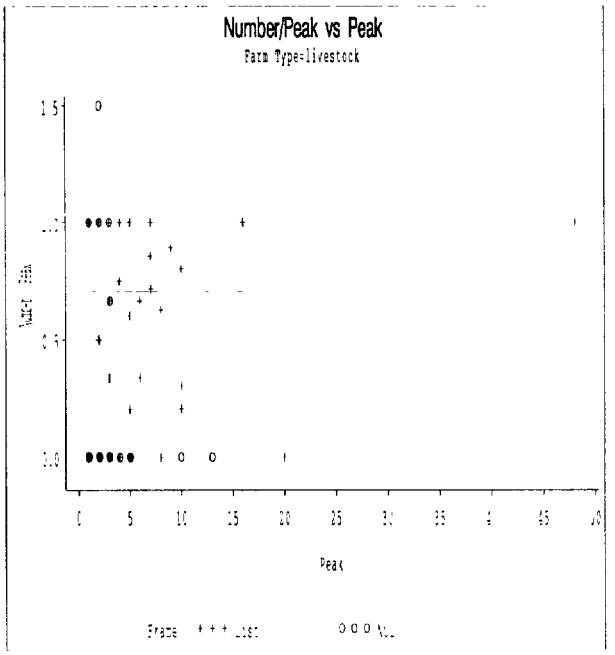


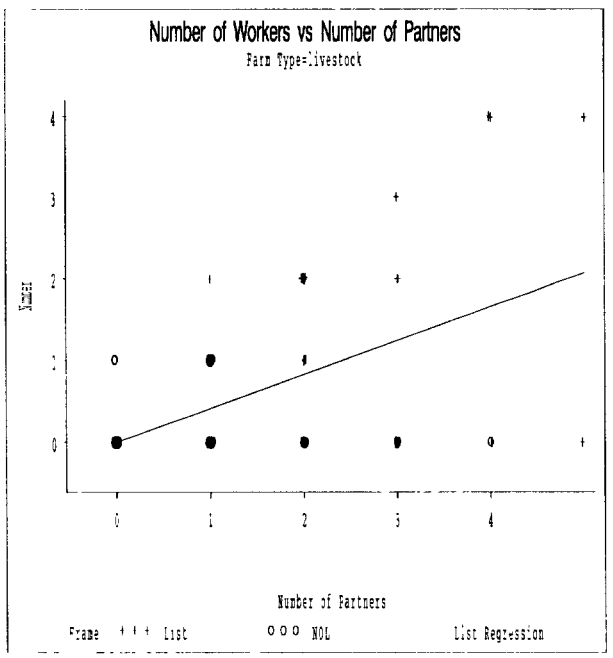
Figure A4: North Plains

Hired Workers





Self-Employed Workers



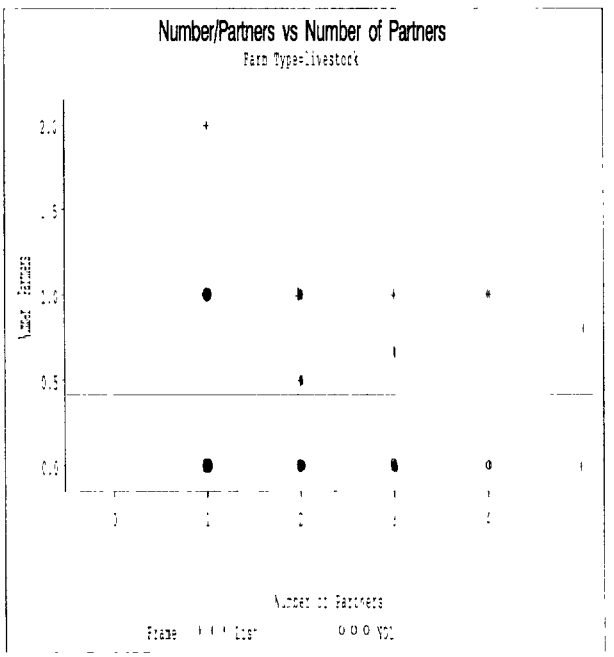


Table A1: **Correlations Between Hired and Peak Number of Workers**

Farm Type	List Correlation	List Sample Size	NOL Correlation	NOL Sample Size
Appalachian I† - July 1995				
Grains	0.82	22	*	0
Tobacco	0.83	173	0.84	17
Oth Crop / Oth Livestock	0.45	21	0.73	4
Vegetables / Dairy	0.99	29	0.84	4
Nursery	0.97	27	*	0
Cotton / Livestock	0.81	74	0.86	21
Poultry / Fruits	0.39	50	*	1
California - July 1995				
Grains	0.25	22	0.85	11
Cotton	0.99	16	0.99	3
Other	0.95	30	0.91	11
Vegetables	0.53	40	0.99	8
Fruits	0.85	205	0.96	27
Nuts	0.69	64	0.40	7
Nursery / Oth Livestock	0.99	44	*	*
Livestock	0.90	45	0.94	14
Poultry / Dairy	0.99	61	0.43	11
Florida - July 1995				
Vegetables / Oth Crop	0.65	39	0.64	6
Fruits / Oth Livestock	0.97	39	0.70	14
Nursery / Poultry	0.92	82	0.99	13
Livestock	0.63	23	0.98	20
North Plains‡ - July 1995				
Grains / Nursery / Poultry / Oth Live	0.34	168	0.86	28
Oth Crop	0.61	16	*	*
Livestock	0.90	90	*	*
Dairy	0.94	22	0.26	20

* The correlation could not be calculated because the sample size was insufficient.
† Appalachian I region is comprised of Virginia and North Carolina.
‡ North Plain region is comprised of Kansas, Nebraska, North Dakota and South Dakota.

Table A2: **Correlations Between Self-Employed Workers and Number of Partners**

Farm Type	List Correlation	List Sample Size	NOL Correlation	NOL Sample Size
Appalachian I† - July 1995				
Grains	0.58	55	*	*
Tobacco / Oth Livestock	0.82	194	0.64	17
Other Crops	-0.66	24	0.5	3
Vegetables	0.14	19	*	*
Fruits	0.64	16	*	*
Nursery	0.11	32	*	*
Livestock / Cotton	0.67	123	0.12	22
Poultry / Oth Livestock	0.54	64	*	*
Dairy	0.07	16	*	*
California - July 1995				
Grains	0.32	31	0.90	11
Cotton	0.52	18	0.50	3
Other Crops	0.82	44	0.96	7
Vegetables / Poultry	0.55	45	0.48	8
Fruits	0.48	263	0.73	20
Nuts	0.55	81	*	*
Nursery	0.56	40	*	*
Livestock / Oth Livestock	0.46	100	0.63	12
Dairy	0.83	57	0.82	4
Florida - July 1995				
Vegetables	-0.23	24	-0.61	5
Fruits / Cotton	-0.03	67	*	*
Nursery / Oth Crop / Oth Live	-0.04	85	*	*
Poultry	-0.12	31	*	*
North Plains‡ - July 1995				
Grains / Nursery / Poultry / Oth Live	0.58	353	0.51	28
Oth Crop	0.85	22	*	*
Livestock	0.49	222	*	*
Dairy	0.26	31	*	*
Christmas Tree	0.50	21	*	*

* The correlation could not be calculated because the sample size was insufficient.

† Appalachian I region is comprised of Virginia and North Carolina.

‡ North Plain region is comprised of Kansas, Nebraska, North Dakota and South Dakota.

APPENDIX B: REGION LEVEL RESULTS

List of Figures

Using the 1992-93, 93-94, 94-95 and 95-96 Quarterly Agricultural Labor Surveys from states in each of the 17 agricultural regions in the United States, estimates of the labor items of interest (total number of workers and hours worked per week for the hired, self-employed and unpaid workers, as well as the wage rates for the hired workers) were compared to the DE at the regional and national levels. The relative mean deviation (R-MD) and the relative root mean squared deviation (R-RMSD) were used for the performance criterion. Table B1 provides a listing of the states included in each of the 17 agricultural regions in the United States.

Table B1. **United States Agricultural Regions**

Region	States Included
Appalachian I	NC,VA
Appalachian II	KY,TN,WV
California	CA
Cornbelt I	IL,IN,OH
Cornbelt II	IA,MO
Delta	AR,LA,MS
Florida	FL
Lake	MI,MN,WI
Mountain I	ID,MT,WY
Mountain II	CO,NV,UT
Mountain III	AZ,NM
Northeast I	CT,ME,MA,NH,NY,RI,VT
Northeast II	DE,MY,NJ,PA
North Plains	KS,NE,ND,SD
Pacific	OR,WA
Southeast	AL,GA,SC
Southern Plains	OK,TX

Figure B1: Relative Mean Deviation from DE

Hired (Total)

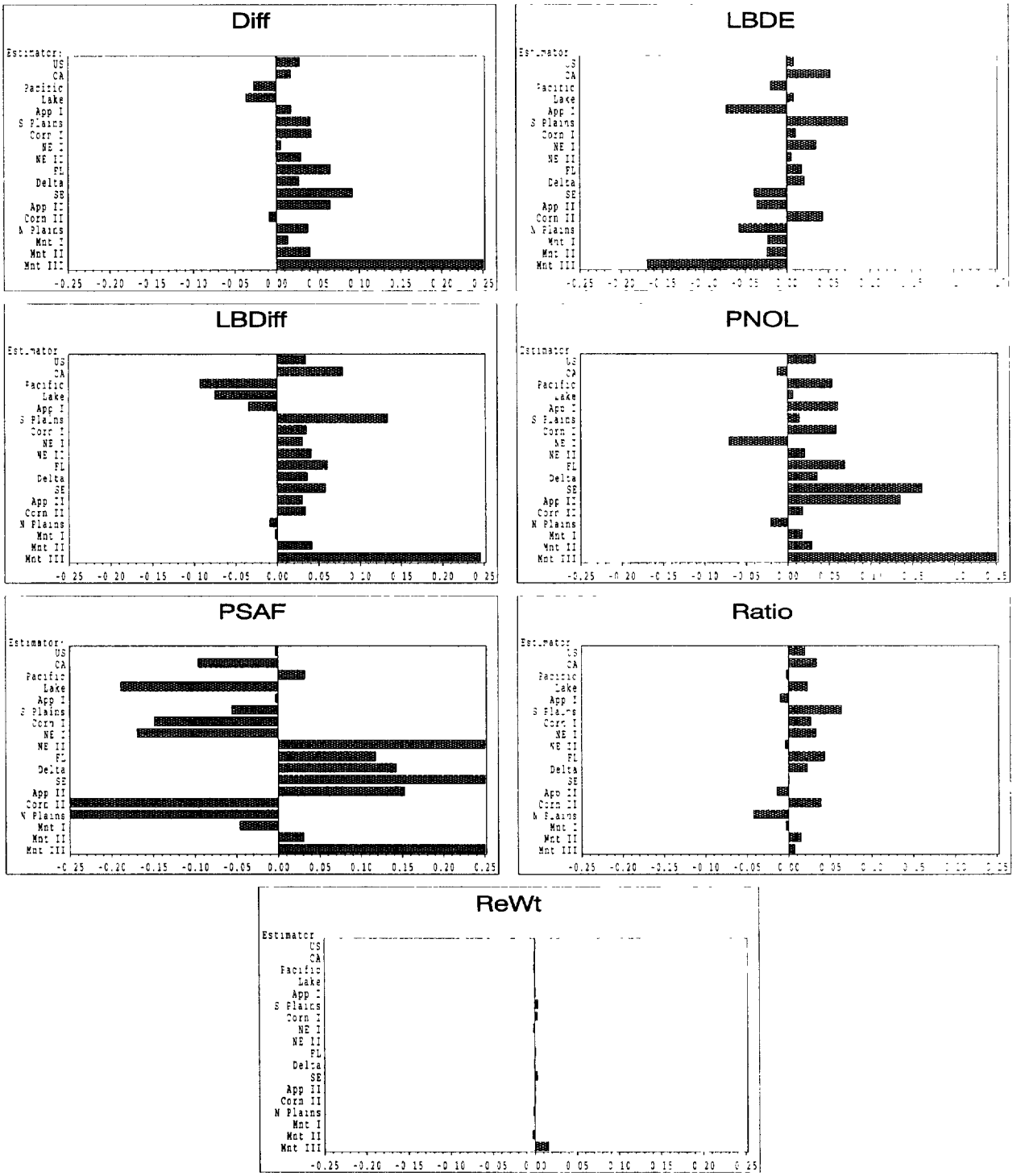


Figure B2: Relative Mean Deviation from DE

Hired (Weekly Hours)

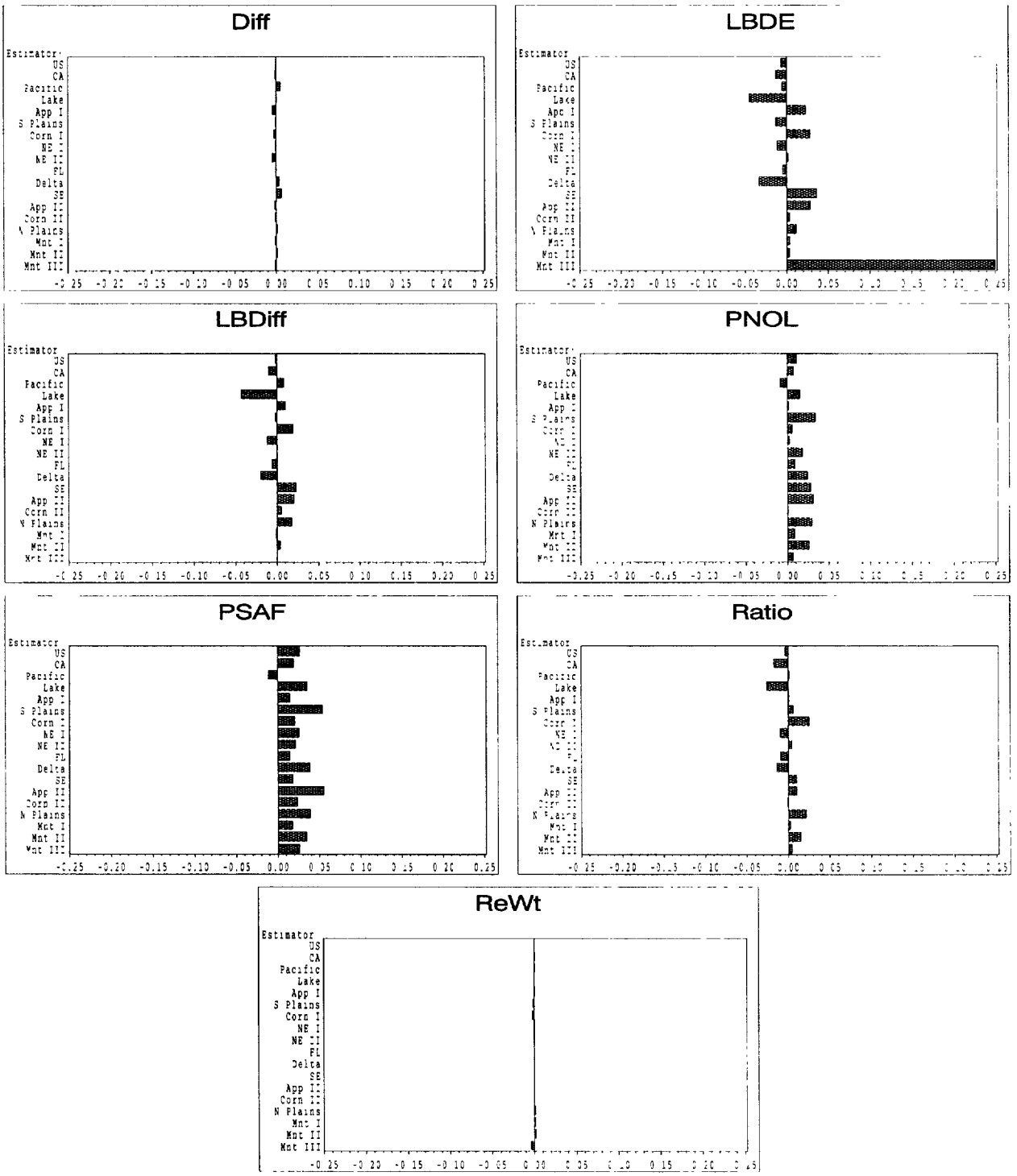


Figure B3: Relative Mean Deviation from DE

Hired (Wage Rates)

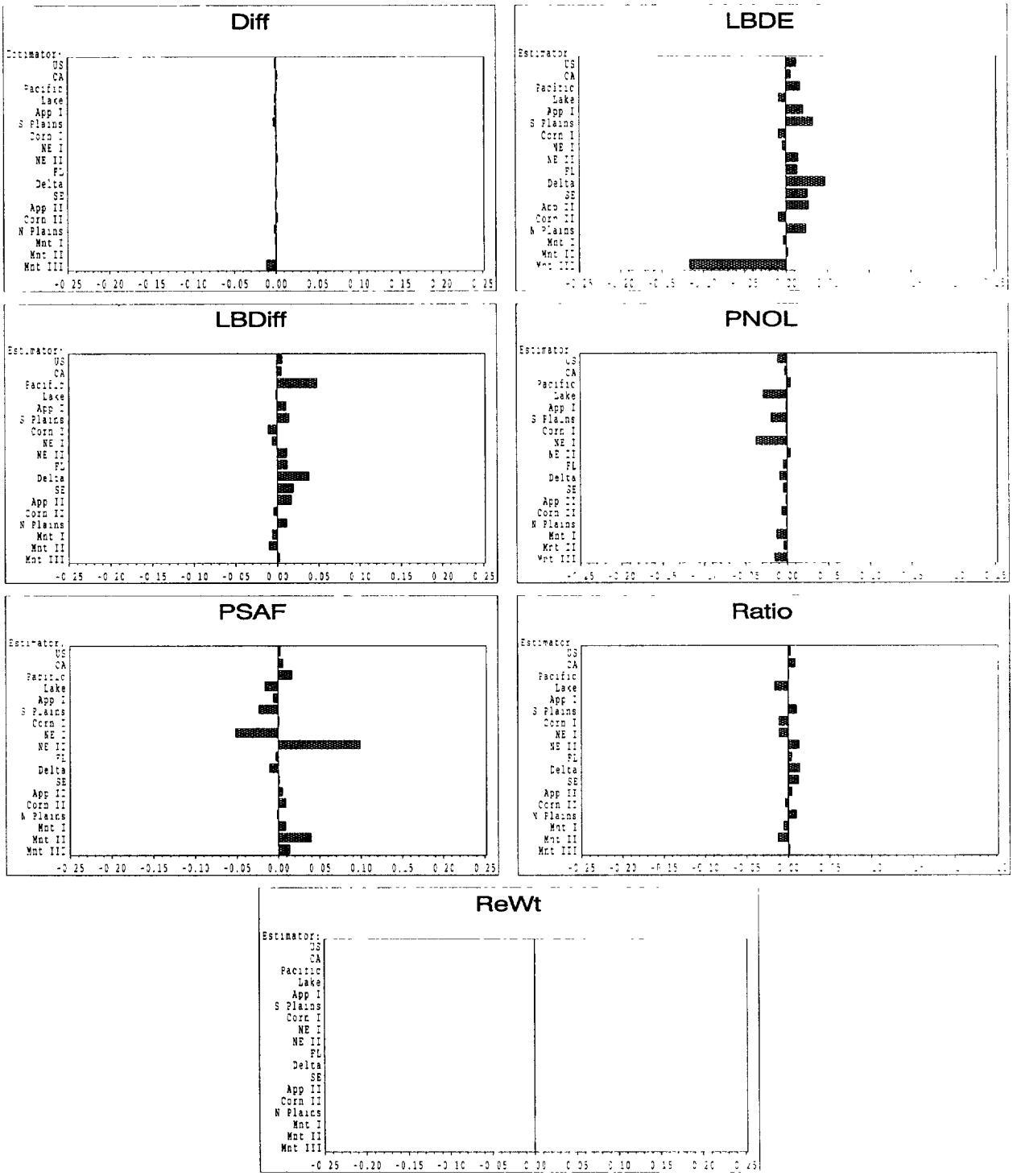


Figure B4: Relative Mean Deviation from DE

Self-Employed (Total)

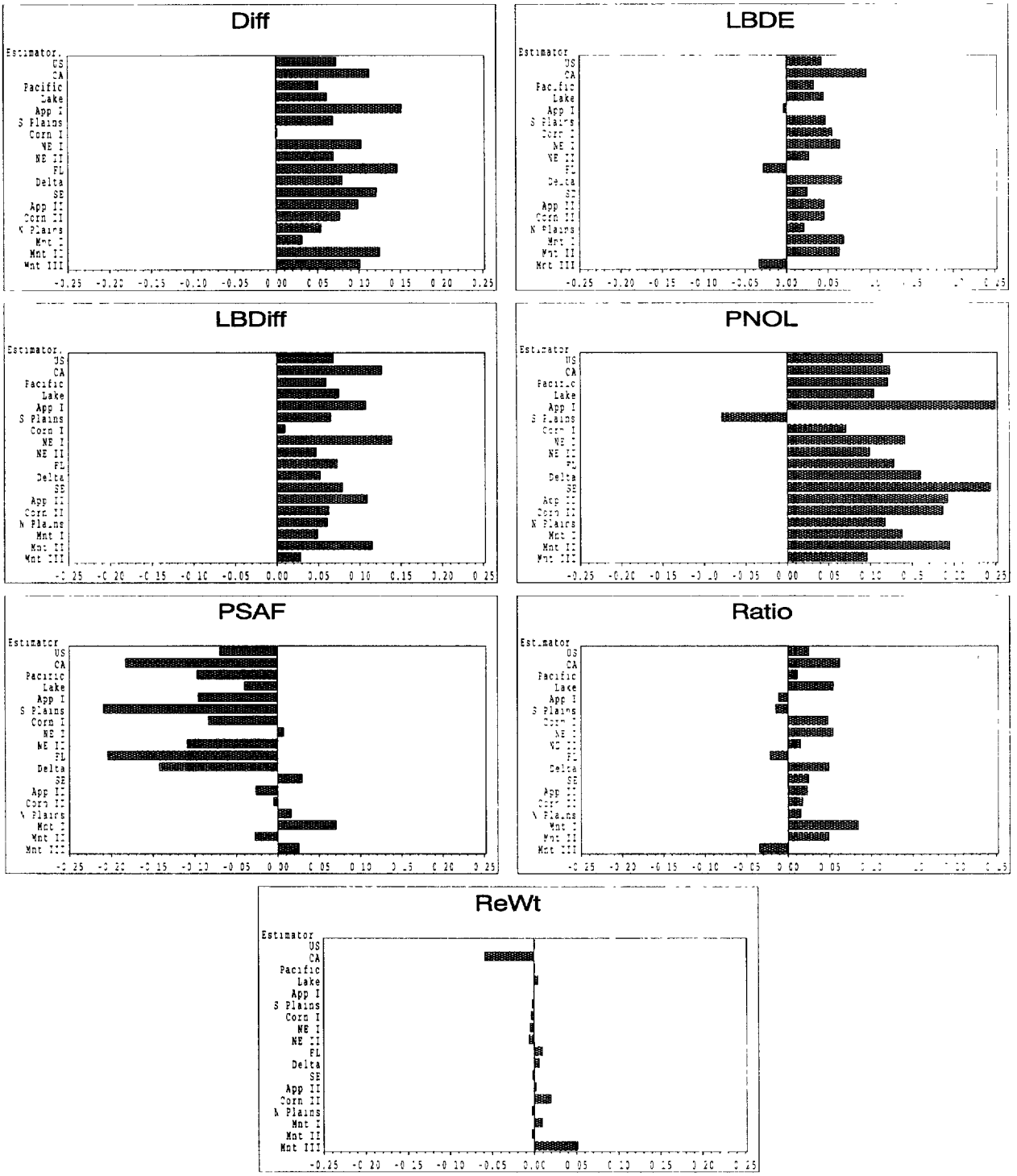


Figure B5: Relative Mean Deviation from DE

Self-Employed (Weekly Hours)

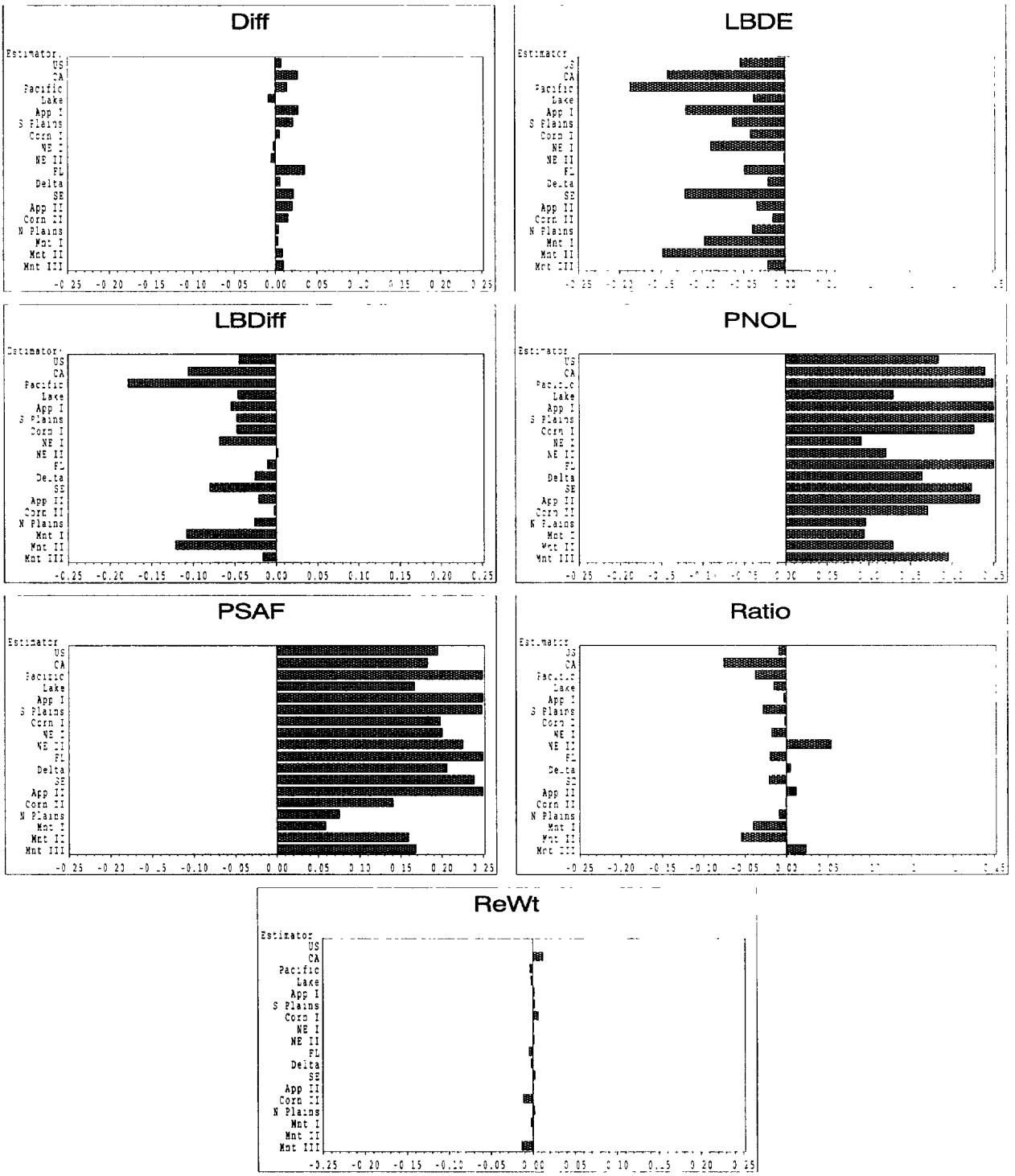


Figure B6: Relative Mean Deviation from DE

Unpaid (Total)

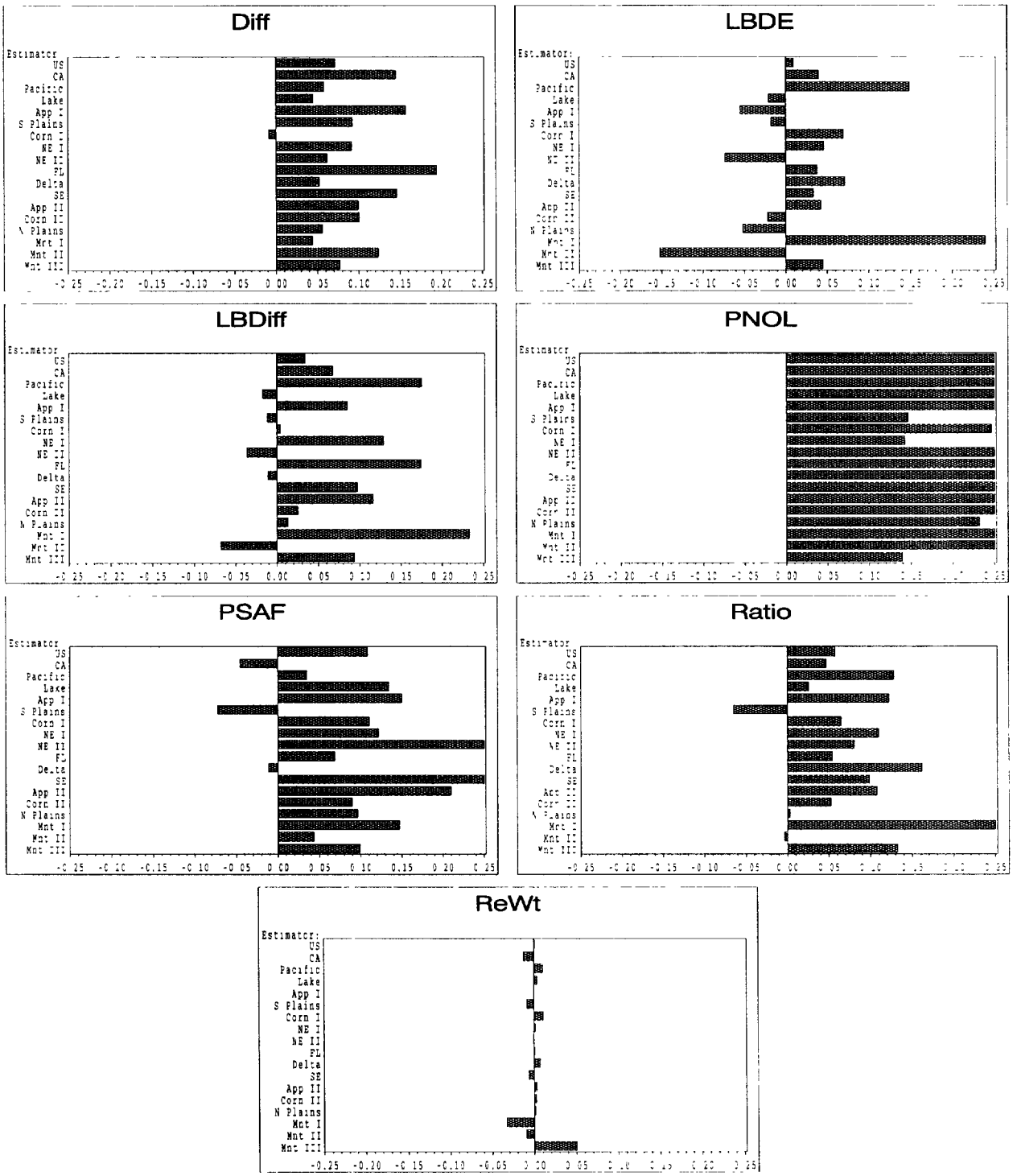


Figure B7: Relative Mean Deviation from DE

Unpaid (Weekly Hours)

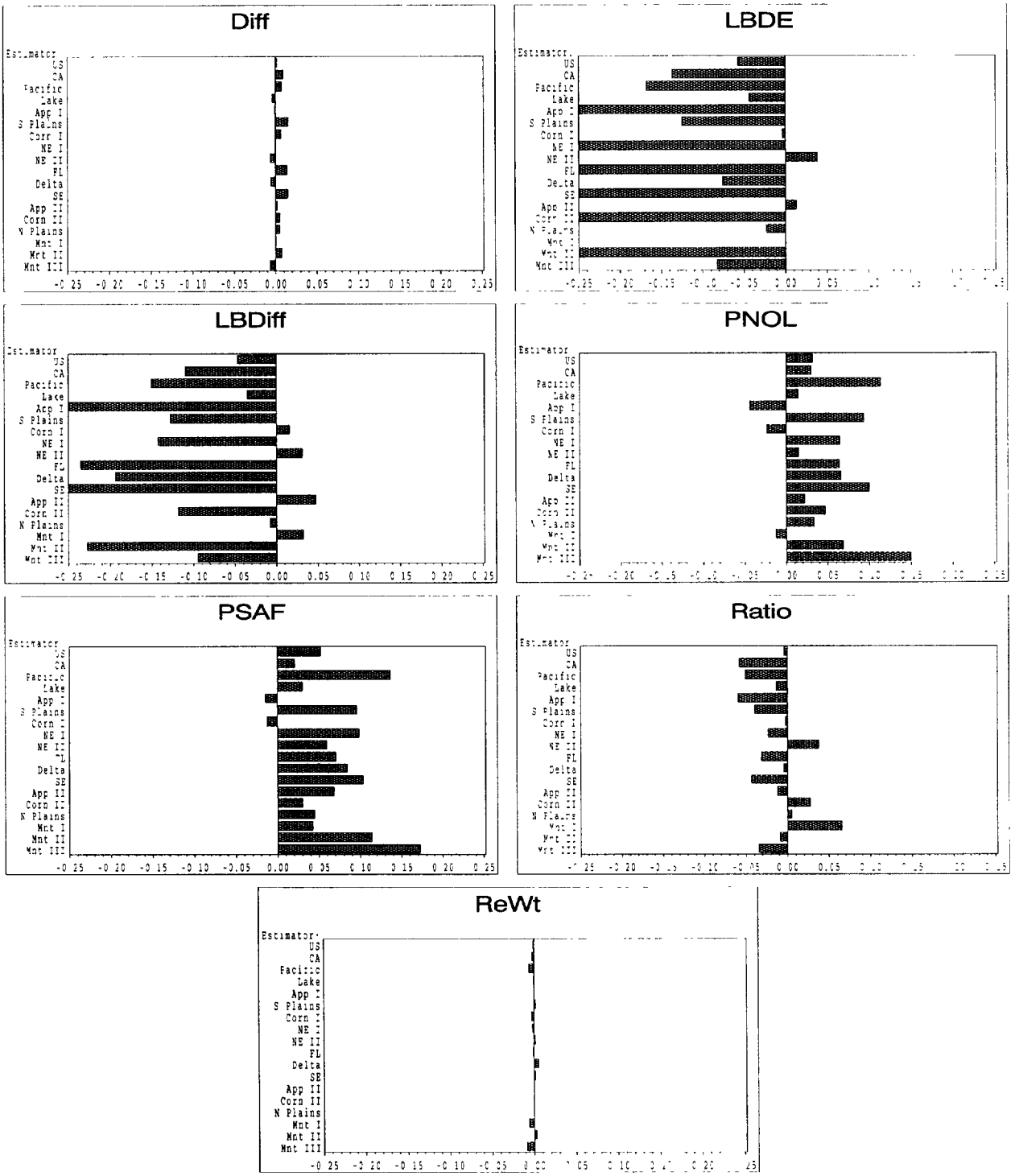


Figure B8: Relative Root Mean Square Deviation from DE

Hired (Total)

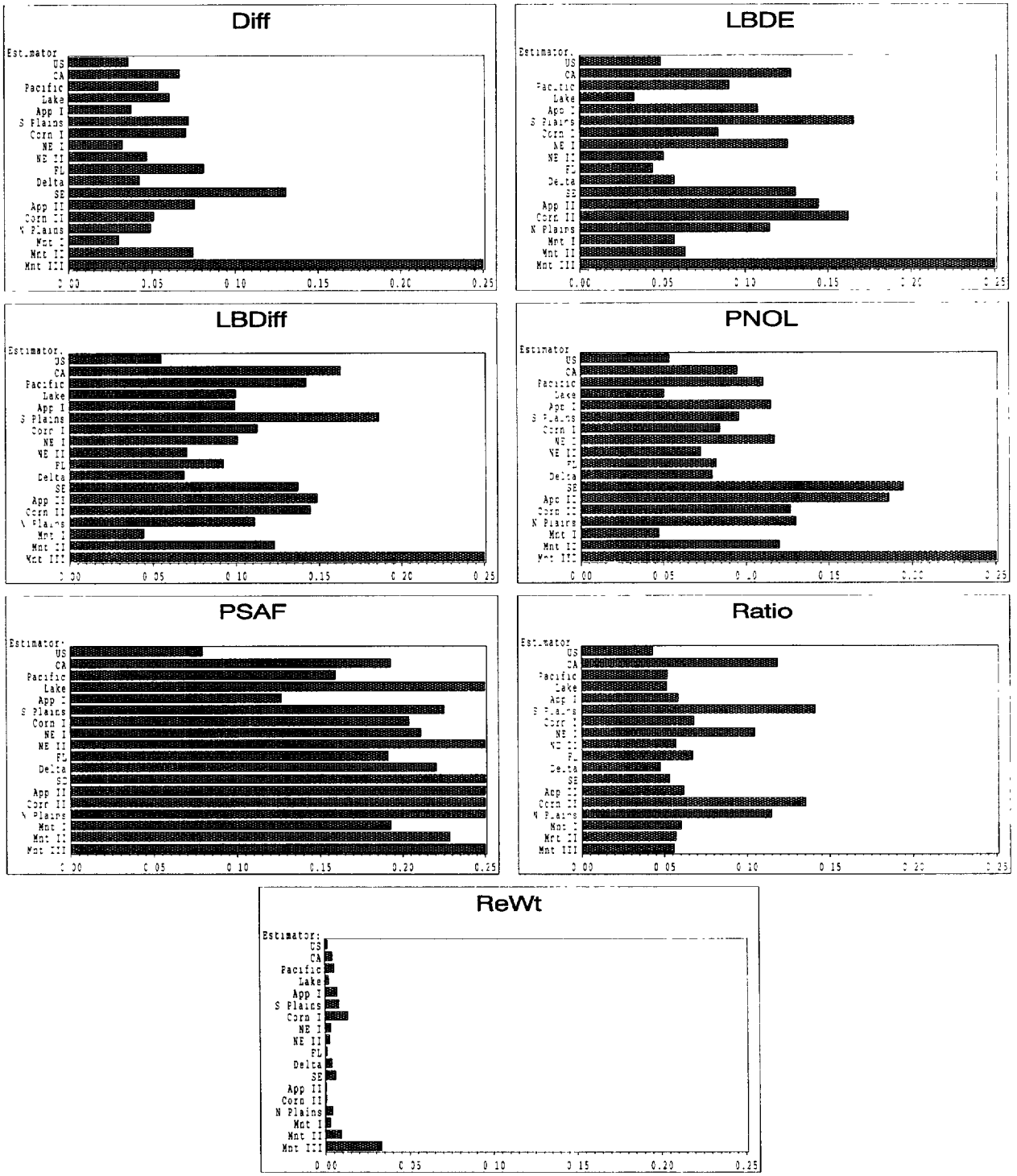


Figure B9: Relative Root Mean Square Deviation from DE

Hired (Weekly Hours)

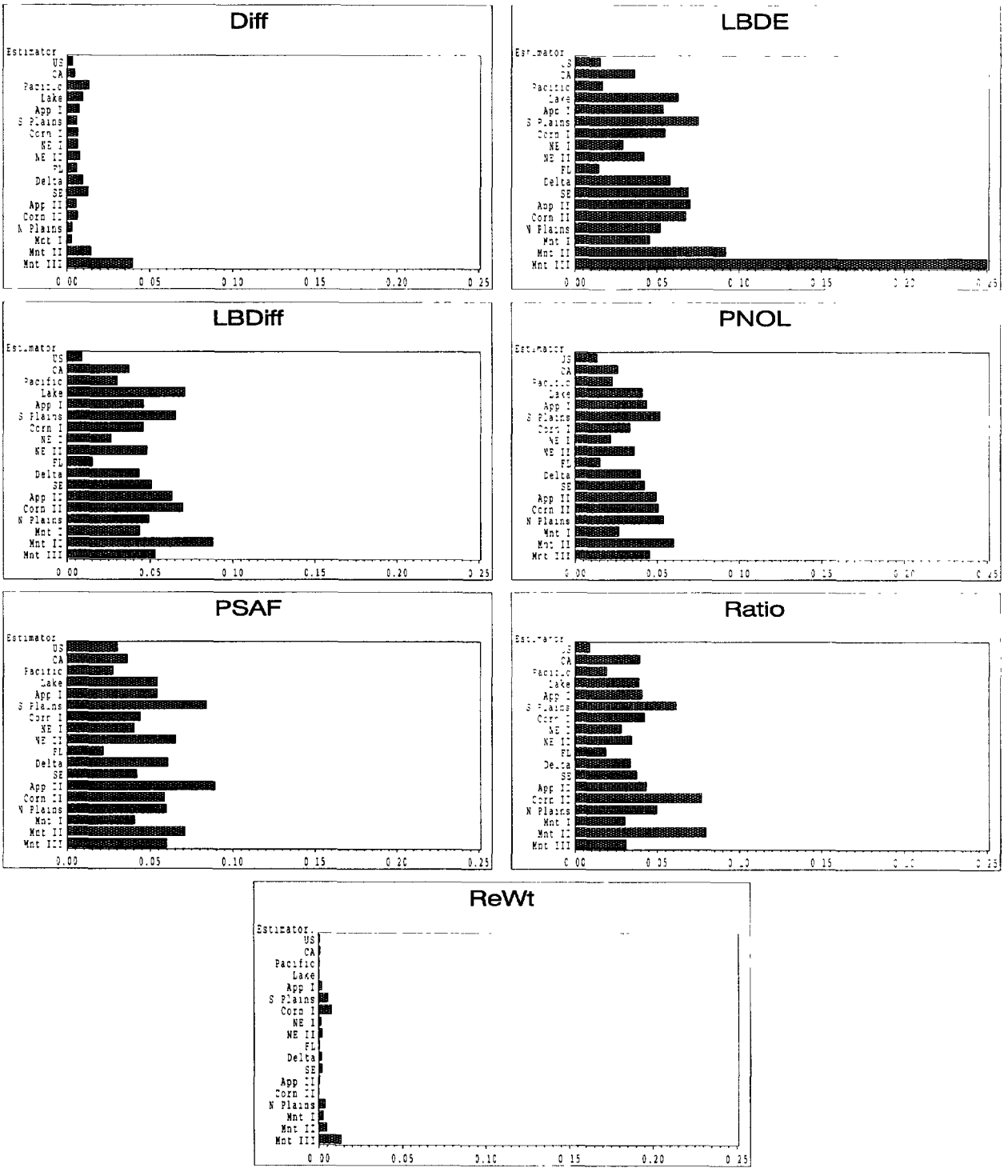


Figure B10: Relative Root Mean Square Deviation from DE

Hired (Wage Rates)

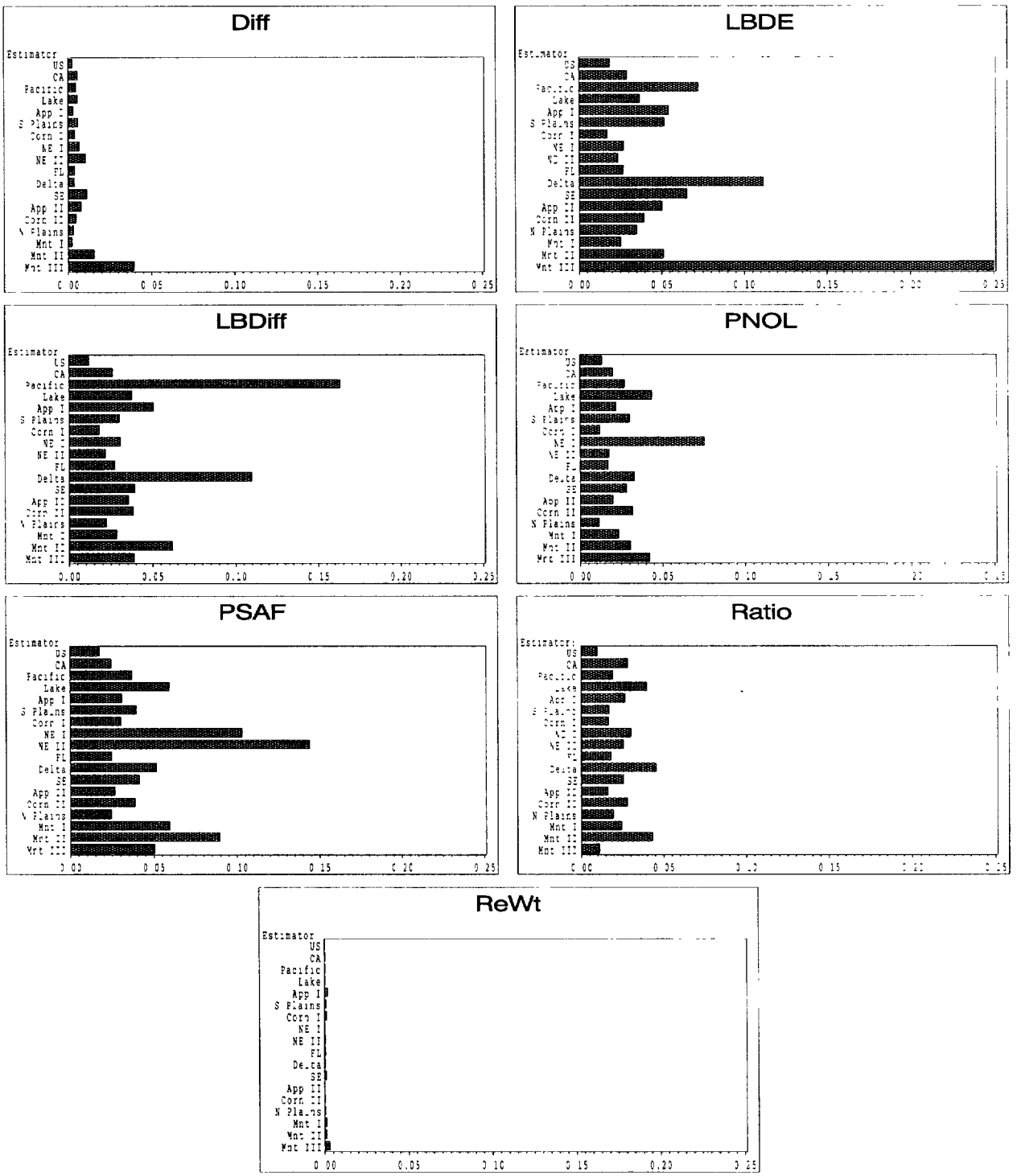


Figure B11: Relative Root Mean Square Deviation from DE

Self-Employed (Total)

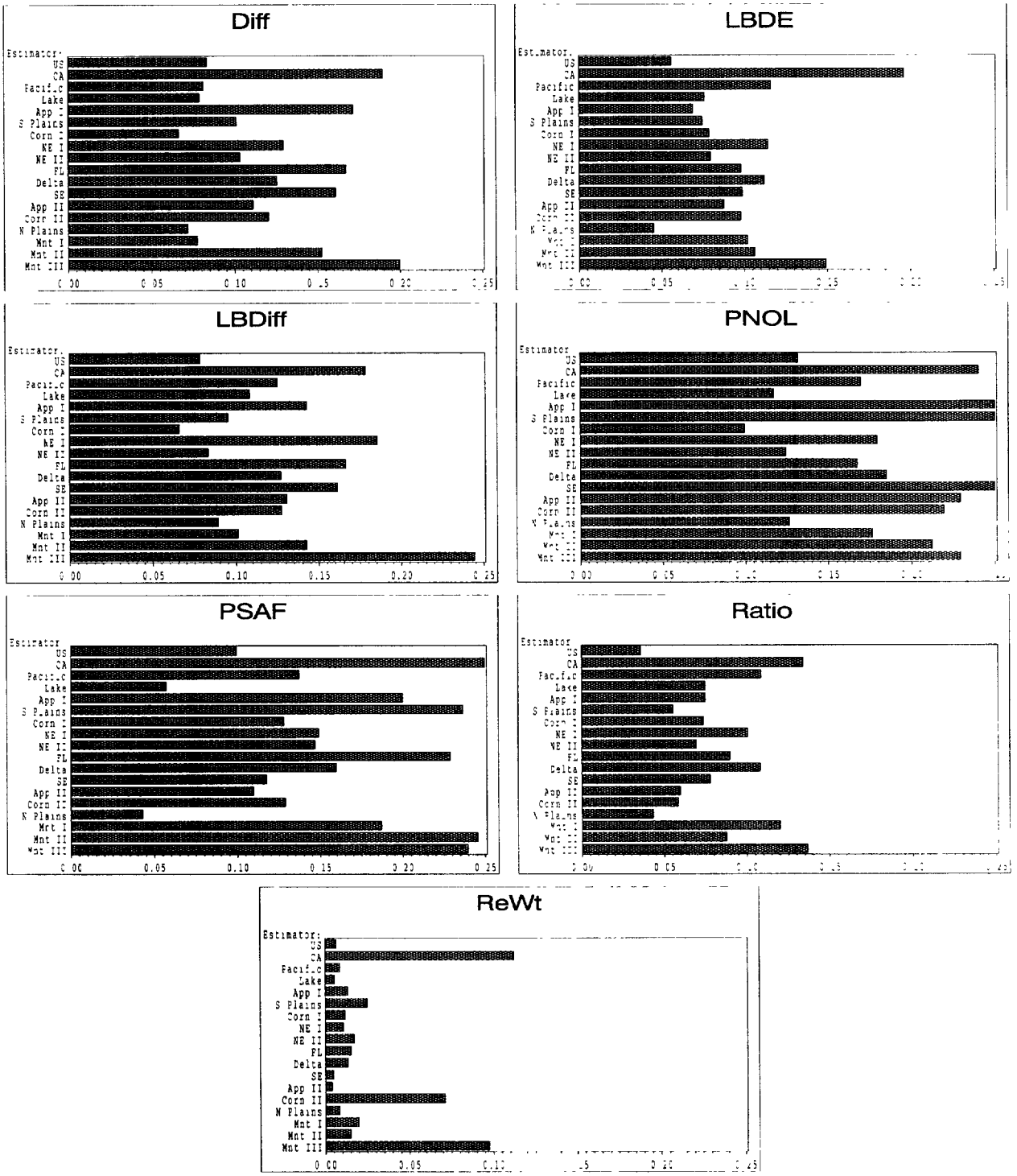


Figure B12: Relative Root Mean Square Deviation from DE

Self-Employed (Weekly Hours)

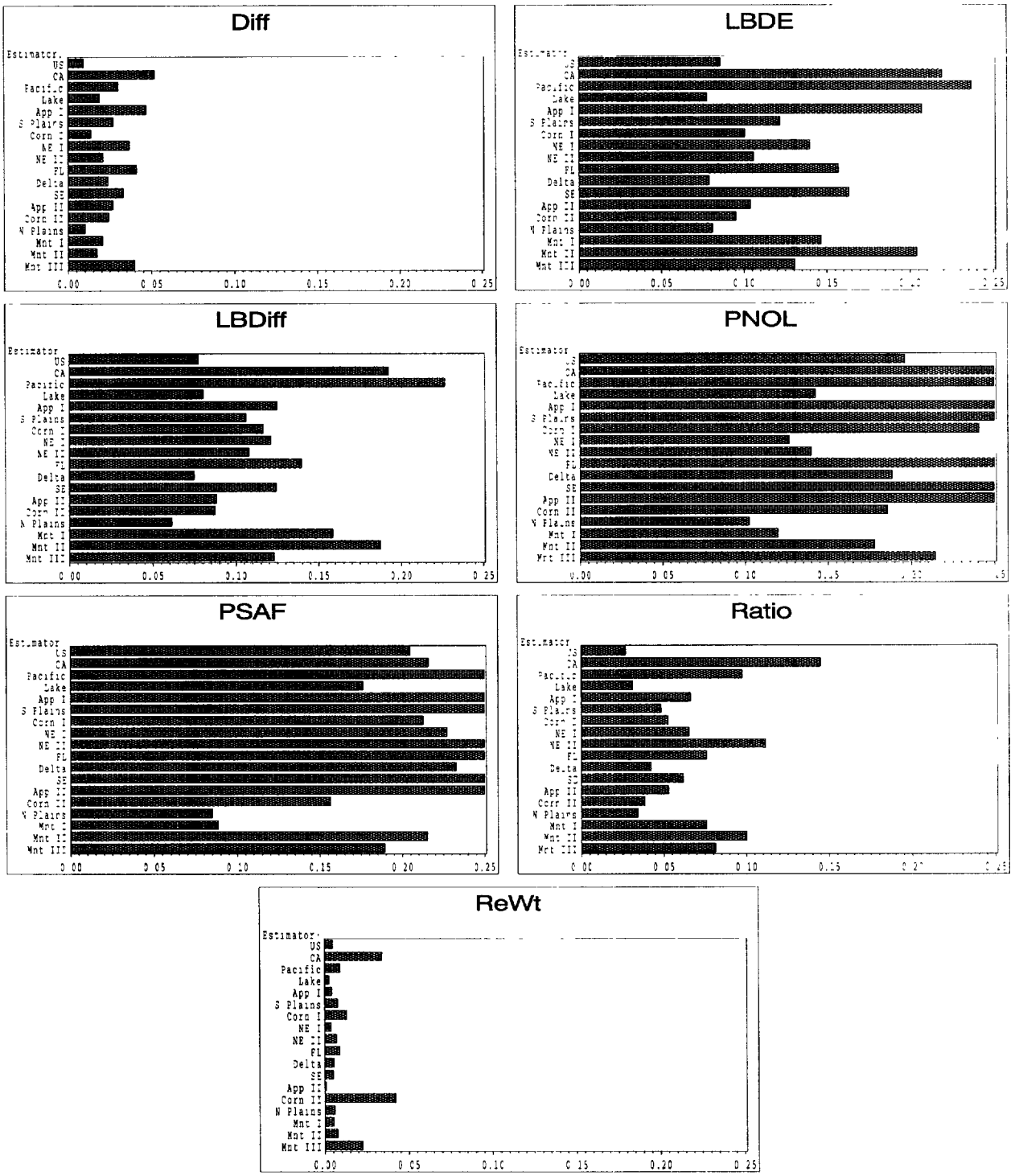


Figure B13: Relative Root Mean Square Deviation from DE

Unpaid (Total)

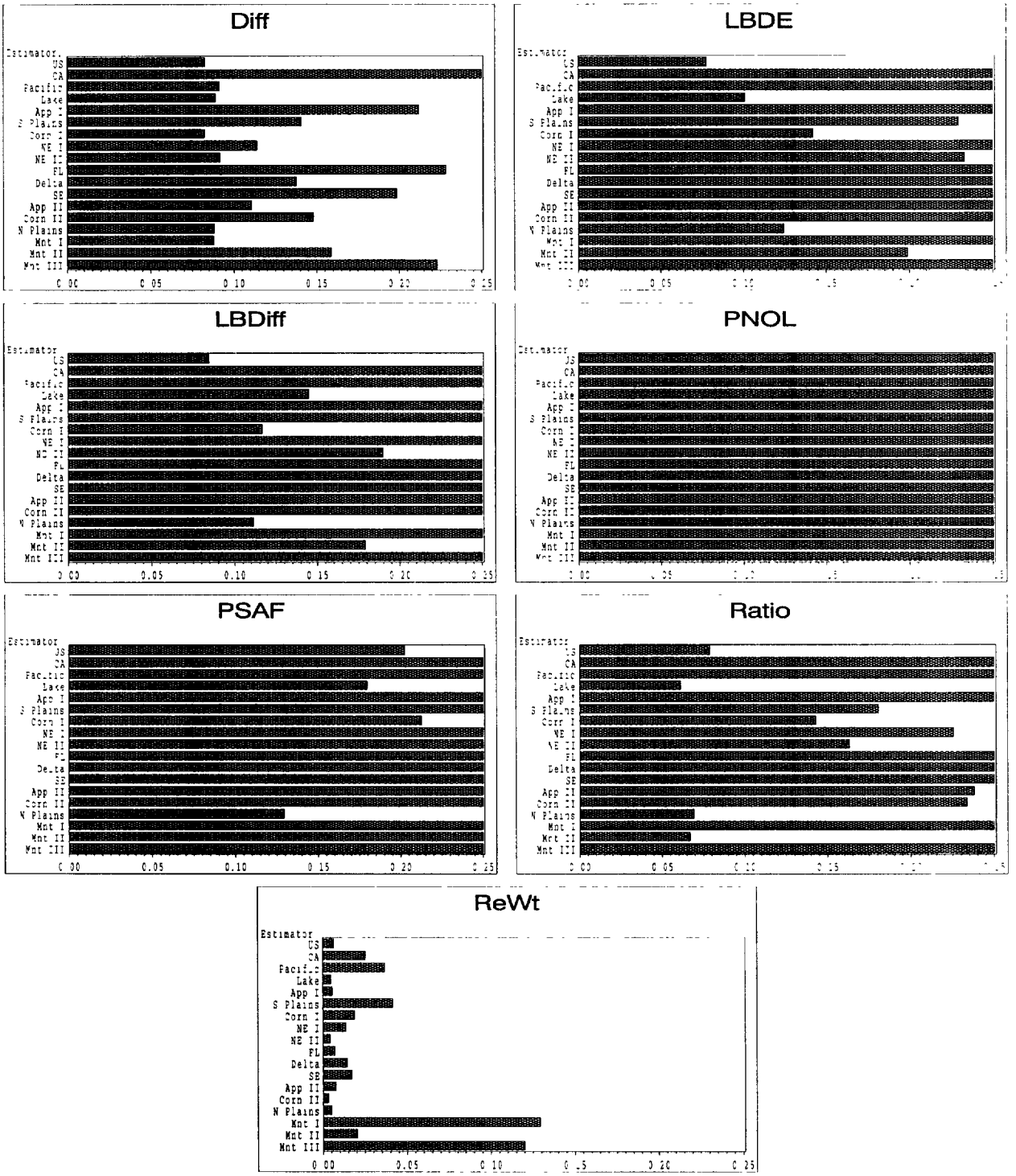
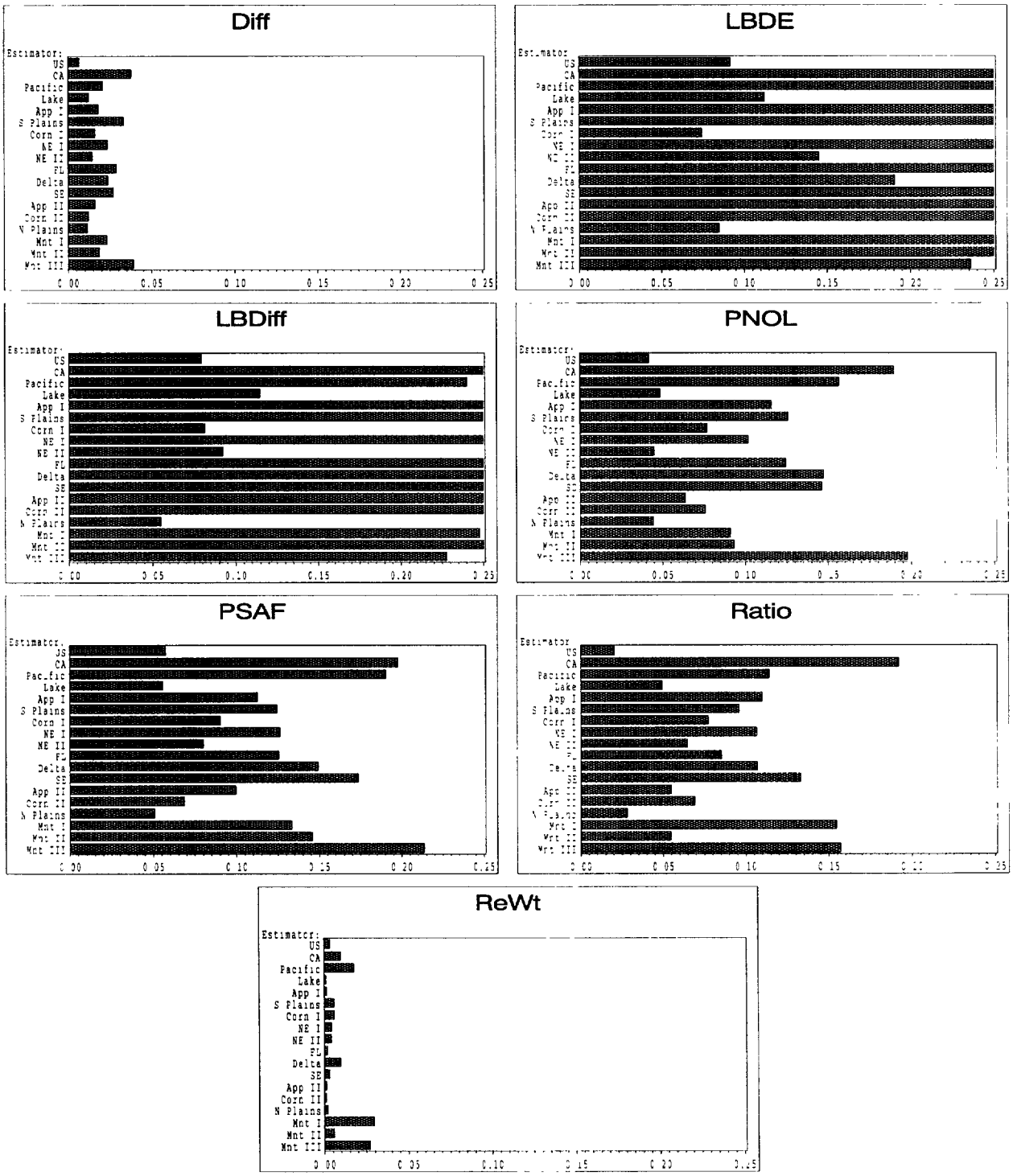


Figure B14: Relative Root Mean Square Deviation from DE

Unpaid (Weekly Hours)



APPENDIX C: JACKKNIFE PROCEDURE

Computations and Evaluations

The closed form variance formulae are not available for most of the alternative estimators that have been studied for the labor survey. Thus, their variance estimates must be obtained by using a resampling procedure, or by using some other procedure such as linearization. The resampling procedure with the most desirable characteristics for our applications is the delete-a-random-group jackknife. In comparison to other procedures, it is easy to understand and apply. For most NASS sampling designs, the delete-a-random-group jackknife procedure should be relatively robust and provide slightly conservative variance estimates. Since it is used to obtain variance estimates for all of the proposed estimators, the relative merits of each estimator can be examined.

Computing delete-a-random-group jackknife variance estimates is conceptually very simple. (1) Divide the sample into R random groups (pseudo-replicates) in such a way that each random group has essentially the same sampling design as the parent sample. (2) Form R jackknife replicated samples by dropping one of the random groups at a time from the parent sample. (3) Compute R sets of jackknife replicate sampling weights, one for each jackknife replicated sample, by treating, in turn, each of the jackknife replicated samples as the sample. (4) Compute R jackknife replicated estimates by carrying out the estimation process once for each jackknife replicated sample and its associated replicated sampling weights. (5) Compute the delete-a-random-group jackknife variance estimate for the estimator involved, which is (R-1)/R times the sum of the squares of the deviations of the R jackknife replicated estimates from their mean. A slightly more conservative approach is to use (R-1)/R times the sum of the square of deviations of the R jackknife replicated estimates from the estimate from the full sample.

In spite of the fact that each step of this procedure is basically straight forward and easy to understand, a measure of subtlety is generally required to properly apply them to a particular design since a number of practical problems normally arise. Therefore each step, as it applies to the quarterly labor survey, is elaborated below.

Division Of The Sample Into R Random Groups.

The first step is to partition the first phase sample into R random groups (pseudo-replicates). If multiple phase sampling is involved, later phase units are associated with the random groups on the basis of their first phase unit. The basic idea is to divide the first phase units of the parent sample into R mutually exclusive groups of equal size in such a way that each group can be thought of as a miniature first phase sample that mimics the first phase sampling of the parent sample.

Thus, the parent sample must be divided, on the basis of its first phase units, into R groups in such a way that the first stage units of the parent sample can be thought of as being drawn in R replicates without replacement. In fact, if one knows that when the first phase

of the parent sample is drawn, the variances are to be estimated by the delete-a-random-group jackknife procedure, then it is generally desirable to form the random groups at that time using an appropriate method of replicated sampling to draw the first phase units. The method used to divide the labor survey sample into random groups is based on the usual technique of forming rotation groups. The method, which is essentially the same for both the list and NOL samples, can be summarized as follows. First, sort the first phase sampling units of parent sample so that when they are systematically assigned to R random groups, each random group will mimic the first phase sampling of the parent sample. Systematically assign the first phase units to the R random groups. Then, associate the later phase units to the random groups according to the location of the first stage units. Since different sorting and assignment techniques are required for the list and area frame samples, each sample is discussed separately.

List: Since the list frame portion of the quarterly labor survey uses a rotating stratified sample with 50% quarter to quarter overlap, these aspects must be duplicated in each of the random groups. This means that the random groups must be formed at the rotation group level to mimic the stratified design or equivalently at the strata level to mimic the rotation groups.

The annual labor list sample is made up of four rotation groups, which we can denote by 1, 2, 3, and 4. The 50% quarter to quarter rotation pattern is obtained by using groups (1 and 2), (2 and 3), (3 and 4), and (4 and 1) for the July, October, January, and April surveys, respectively.

With this background, the method of dividing the list frame portion of the parent sample into R random groups is described by the following three steps.

1. Form a set containing all records that are used in any of the four quarters – records are listed once and only once. Each record should include an id number, rotation group number, stratum code, and a randomly assigned number.

2. Sort this set first by rotation group, then by stratum within each rotation group, and finally by the randomly assigned number within rotation group and stratum. (If the parent sample was selected systematically, the final sort should be according to the systematic order of selection.)

3. Systematically assign the sorted records to the R random groups by placing records 1, R+1, 2R+1, 3R+1, ... in random group 1; placing records 2, R+2, 2R+2, 3R+2, ... in random group 2; etc.; placing records R, 2R, 3R, ... in random group R.

Area: Since some list-only estimators use auxiliary information from both the June area frame sample segments that are used in labor survey (40%) and the June area frame sample segments that are not used in the labor survey (60%), this so called "60-40" split of the June segments must be reflected in each of the random groups. This implies that the random groups must be formed separately within each part of the original 60-40 split of the June

segments. That is, the groups should be formed separately within the group of June sample segments that are used in the labor survey and separately in the group of June sample segments that are not used in the labor survey.

The area frame portion of the quarterly labor survey sample is drawn in two phases. In the first phase a stratified design is used to draw a set of area frame segments. Then, a list of all the operations that are contained in the first phase sample segments is developed and stratified according to various size criterion. In the second phase, a stratified sample of operations is selected. (In forming the random groups, it is important to remember how the area frame is stratified and sampled. The area frame is first divided into land-use strata. Then, each land-use stratum is serpentinely divided into the substrata which become the sampling strata. The area frame sample is selected by taking a simple random sample of segments from each substratum. However, the selection probabilities are essentially the same for each substratum of a land-use stratum.)

The method used to divide the list portion of the sample into random groups is easily adapted to divide the first phase segments of the June area sample into R random groups. Within the area frame portion of the labor survey, the R random groups of ultimate interest are formed by assigning the second phase units (operations) to the R random groups according to the location of the associated first phase units (segments). The method is completely specified by the following four steps.

1. Form a set containing one record for each June area frame sample segment. Include on each record a code, say "uselabor", that distinguishes the records used in the area frame portion of the labor survey from those not used in the labor survey – each segment is listed once and only once. Each record should include the segment id number, uselabor code, stratum code, substratum code, and a randomly assigned number.

2. Sort this set first by uselabor, second by stratum within uselabor, third by substratum within uselabor and stratum, and finally by the random assigned number within uselabor, stratum, and substratum.

3. Systematically assign the sorted segments to the R random groups by placing segments 1, R+1, 2R+1, 3R+1, ... in random group 1; placing segments 2, R+2, 2R+2, 3R+2, ... in random group 2; etc.; placing segments R, 2R, 3R, ... in random group R.

4. Assign all the operations from the first phase sample of sample segments to R random groups on the basis of the assignment of their first phase segments, which means that an operation coming from segment kR+i is assigned to random group i. All the operations contained in the sample segments must be included in the random groups because they will be needed later, when the jackknife replicate weights are computed for the second phase, in determining the jackknife replicate's second phase population and sample size. This implies that the operation records must carry a data item indicating whether or not the operation is included in the sample.

Determination of the R jackknife replicated samples.

The R jackknife replicates are formed by deleting from the initial sample one of the random groups at a time. That is, the rth jackknife replicate for a particular quarter contains all the records initially sampled for that quarter that are not in that quarter's rth random group. For example, the rth jackknife replicate in the October quarter would contain: 1) the list frame sampled records that are in rotation groups 2 and 3 which are not in the rotation groups 2 and 3 portion of random group r and 2) the June area frame sample records that are not in rotation group r.

Derivation Of The R Sets Of Jackknife Replicate Weights.

The basic idea is to attach R sets of sampling weights to the records of the initial quarterly samples (and to the records of the June area frame samples) in such a way that the rth set of weights reflect the sampling weights that the records would have received had the initial sample been the rth jackknife replicated sample. The rth set of weights is called rth jackknife replicate weights, since they are the sampling weights for the rth jackknife replicate. The rth set of jackknife replicate weights to be used with the data collected in the labor survey from the records in the list frame portion of a quarterly labor survey is given by:

$$w_i(r) = \begin{cases} 0 & \text{if list record } i \text{ is not in jackknife replicate } r. \\ w_i(n_h/n_h(r)) & \text{if record } i \text{ is in jackknife replicate } r, \end{cases}$$

where $w_i(n_h/n_h(r)) = (N_h/n_h)(n_h/n_h(r)) = N_h/n_h(r)$, and

- w_i is the initial sample weight for record i of list stratum h,
- N_h is the population size of list stratum h,
- n_h is the initial effective sample size for list stratum h, and
- $n_h(r)$ is the jackknife replicate r effective sample size for list stratum h.

The rth set of jackknife replicate weights to be used with the data collected in the labor survey from the records in the area frame portion of a quarterly labor survey is given by:

$$w_i(r) = \begin{cases} 0 & \text{if area record } i\text{: 1) is not in jackknife replicate } r \text{ or 2) is in jackknife replicate } r \text{ but not in the second phase sample.} \\ w_i(\frac{n_h}{n_h(r)})(\frac{m_k}{m_k(r)})(\frac{M_k(r)}{M_k}) & \text{if record } i \text{ is in jackknife replicate } r \text{ of area frame sample of operations for the quarter} \end{cases}$$

where

$$w_i\left(\frac{n_h}{n_h(r)}\right)\left(\frac{m_k}{m_k(r)}\right)\left(\frac{M_k(r)}{M_k}\right) = \left(\frac{N_h}{n_h}\right)\left(\frac{M_k}{m_k}\right)\left(\frac{n_h}{n_h(r)}\right)\left(\frac{m_k}{m_k(r)}\right)\left(\frac{M_k(r)}{M_k}\right)$$
$$= \left(\frac{N_h}{n_h(r)}\right)\left(\frac{M_k(r)}{m_k(r)}\right),$$

and

- w_i is the initial sample weight for record i which we assume came from a segment of first phase stratum h and second phase stratum k,
- h is the first phase population size for first phase stratum h,
- n_h is the initial sample size for first phase stratum h,
- $n_h(r)$ is the jackknife replicate r sample size for first phase stratum h.
- M_k is the initial population size for second phase stratum k,
- $M_k(r)$ is the jackknife replicate r population size for second phase stratum k,
- m_k is the initial second phase sample size for second phase stratum k, and
- $m_k(r)$ is the jackknife replicate r sample size for second phase stratum k.

The rth set of jackknife replicate weights to be used with the area frame data collected in the June Agriculture Survey is given by: w_i(r)= 0, if area record i is not in jackknife replicate r. If record i is in jackknife replicate r, then w_i(r)= $w_i(n_h/n_h(r)) = (N_h/n_h)(n_h/n_h(r)) = (N_h/n_h(r))$,

where

- w_i is the initial sample weight for record i which we assume came from a segment of stratum h,
- N_h is the population size for stratum h,
- n_h is the initial sample size for stratum h in June,
- $n_h(r)$ is the jackknife replicate r sample size for stratum h.

Computation Of The R Jackknife Replicated Estimates.

The basic idea is to obtain R jackknife replicated estimates that each completely duplicates the sample and estimation process. For any labor survey estimator T (either list-only or multiple frame) the set of R jackknife replicated estimates, $T_{\text{R}-1}$ (1), $T_{\text{R}-1}$ (2), ..., $T_{\text{R}-1}$

(R), is obtained by applying the estimation process R times, once to each jackknife replicated samples and its associated set of replicated weights.

Computation Of The Jackknife Variance, Bias, and MSE Estimates.

The delete-a-random-group jackknife estimate of the variance of T is obtained by multiplying the sum of squares of deviations of the R jackknife replicated estimates, $T_{\mathrm{R}-1}$ (1), $T_{\mathrm{R}-1}$ (2), ..., $T_{\mathrm{R}-1}$ (R), from their mean by (R-1)/R, which gives:

$$v_1 = ((\mathrm{R}-1)/\mathrm{R}) \sum_{r=1}^{\mathrm{R}} (T_{\mathrm{R}-1}(r) - \bar{T}_{\mathrm{R}-1}(\cdot))^2$$

where

$$\bar{T}_{\mathrm{R}-1}(\cdot) = (1/\mathrm{R}) \sum_{r=1}^{\mathrm{R}} (T_{\mathrm{R}-1}(r)).$$

The delete-a-random-group jackknife estimate of the bias of T is obtained by multiplying the difference between the average of the R jackknife replicated estimates and the full sample estimate by (R-1), which gives:

$$b_1 = (R-1)(\bar{\mathrm{T}}_{n-1,(\cdot)} - T)$$

where T is the estimate obtained from the full sample.

Using the variance and bias estimates above a delete-a-random-group jackknife estimate of the mean square error is obtained by adding the square of the bias estimate to the variance estimate, which gives:

$$\mathrm{mse}_1 = v_1 + (b_1)^2.$$

A slightly more conservative estimate of the variance of T that is often used is given by:

$$v_2 = ((\mathrm{R}-1)/\mathrm{R}) \sum_{r=1}^{\mathrm{R}} (T_{\mathrm{R}-1}(r) - T)^2$$

where T is the estimate obtained from the full sample.

Some Properties Of The Jackknife Variance And Bias Estimates

In most sampling literature, the jackknife estimates are usually expressed in terms of the jackknife replicates in a manner similar to what was done above. However, in most other areas of statistics, these very same estimates are almost always expressed in terms of the jackknife pseudo values. Since this often leads to confusion, the relationship between these two way of expressing the jackknife estimates, along with the relationship between the variance estimates v_1 and v_2, are outlined below.

Let $\mathrm{T}_n = \mathrm{T}_n(x_1, x_2, \ldots, x_n)$ be an estimator of some unknown parameter θ based on the entire sample $x_1, x_2, \ldots, x_n$ and $\mathrm{T}_{n-1,(\alpha)} = \mathrm{T}_{n-1}(x_1, \ldots, x_{\alpha-1}, x_{\alpha+1}, \ldots, x_n)$ be the same statistic based on the $(n-1)$ observations that do not include x_α. Then Quenouille's jackknife estimator of the bias of T_n, which is defined as $\mathrm{E}(\mathrm{T}_n) - \theta$, is given by

$$b_{\text{jack}} = (n-1)(\bar{\mathrm{T}}_{n-1,(\cdot)} - \mathrm{T}_n)$$

where

$$\bar{\mathrm{T}}_{n-1,(\cdot)} = \frac{1}{n}\sum_{\alpha=1}^{n} \mathrm{T}_{n-1,(\alpha)}.$$

Subtracting b_{jack} from the full estimator gives Quenouille's bias reduced jackknife estimator of θ,

$$\mathrm{T}_{\text{jack}} = n\mathrm{T}_n - (n-1)\bar{\mathrm{T}}_{n-1,(\cdot)}$$

which, by the way, always has a higher variance than T_n. In terms of Tukey's "pseudo values",

$$\tilde{\mathrm{T}}_{n-1,(\alpha)} = n\mathrm{T}_n - (n-1)\mathrm{T}_{n-1,(\alpha)},$$

Quenouille's bias reduced estimator is given by

$$\mathrm{T}_{\text{jack}} = \frac{1}{n}\sum_{\alpha=1}^{n} \tilde{\mathrm{T}}_{n-1,(\alpha)} = \frac{1}{n}\sum_{\alpha=1}^{n}(n\mathrm{T}_n - (n-1)\mathrm{T}_{n-1,(\alpha)}).$$

and Quenouille's bias estimator is given by

$$b_{\text{jack}} = \mathrm{T}_n - \mathrm{T}_{\text{jack}} = \frac{1}{n}\sum_{\alpha=1}^{n}(\mathrm{T}_n - \tilde{\mathrm{T}}_{n-1,(\alpha)}).$$

Treating the "pseudo values" as *iid* with variance the same as $\sqrt{n}\mathrm{T}_n$ leads to the jackknife variance estimator v_1:

$$\begin{aligned} v_1 &= \frac{1}{n(n-1)}\sum_{\alpha=1}^{n}(\tilde{\mathrm{T}}_{n-1,(\alpha)} - \frac{1}{n}\sum_{\alpha_1=1}^{n}\tilde{\mathrm{T}}_{n-1,(\alpha_1)})^2 \\ &= \frac{1}{n(n-1)}\sum_{\alpha=1}^{n}(n\mathrm{T}_n - (n-1)\mathrm{T}_{n-1,(\alpha)} - \frac{1}{n}\sum_{\alpha_1=1}^{n}(n\mathrm{T}_n - (n-1)\mathrm{T}_{n-1,(\alpha_1)}))^2 \\ &= \frac{n-1}{n}\sum_{\alpha=1}^{n}(\mathrm{T}_{n-1,(\alpha)} - \frac{1}{n}\sum_{\alpha_1=1}^{n}\mathrm{T}_{n-1,(\alpha_1)})^2 \\ &= \frac{n-1}{n}\sum_{\alpha=1}^{n}(\mathrm{T}_{n-1,(\alpha)} - \bar{\mathrm{T}}_{n-1,(\cdot)})^2. \end{aligned}$$

Similarly v_2 can also be derived from the "pseudo value" as follows:

$$\begin{aligned} v_2 &= \frac{1}{n(n-1)}\sum_{\alpha=1}^{n}(\tilde{\mathrm{T}}_{n-1,(\alpha)} - \mathrm{T}_n)^2 \\ &= \frac{1}{n(n-1)}\sum_{\alpha=1}^{n}(n\mathrm{T}_n - (n-1)\mathrm{T}_{n-1,(\alpha)} - \mathrm{T}_n)^2 \\ &= \frac{1}{n(n-1)}\sum_{\alpha=1}^{n}((n-1)(\mathrm{T}_n - \mathrm{T}_{n-1,(\alpha)}))^2 \\ &= \frac{n-1}{n}\sum_{\alpha=1}^{n}(\mathrm{T}_{n-1,(\alpha)} - \mathrm{T}_n)^2. \end{aligned}$$

This shows that the jackknife estimators v_1 and v_2 have very similar forms when expressed using either the "pseudo value" or jackknife replicate approach. In fact, v_2 can be expressed in terms of v_1 and b_{jack} as:

$$\begin{aligned} v_2 &= \frac{1}{n(n-1)}\sum_{\alpha=1}^{n}(\tilde{\mathrm{T}}_{n-1,(\alpha)} - \mathrm{T}_n)^2 \\ &= \frac{1}{n(n-1)}\sum_{\alpha=1}^{n}\left\{\left[\tilde{\mathrm{T}}_{n-1,(\alpha)} - (\frac{1}{n}\sum_{\alpha_1=1}^{n}\tilde{\mathrm{T}}_{n-1,(\alpha_1)})\right] + \left[(\frac{1}{n}\sum_{\alpha_1=1}^{n}\tilde{\mathrm{T}}_{n-1,(\alpha_1)}) - \mathrm{T}_n\right]\right\}^2 \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n(n-1)} \sum_{\alpha=1}^{n} \left[(\tilde{\mathrm{T}}_{n-1,(\alpha)} - \mathrm{T}_{\mathrm{jack}}) + (\mathrm{T}_{\mathrm{jack}} - \mathrm{T}_n) \right]^2 \\
&= \frac{1}{n(n-1)} \sum_{\alpha=1}^{n} \left[(\tilde{\mathrm{T}}_{n-1,(\alpha)} - \mathrm{T}_{\mathrm{jack}})^2 + 2(\tilde{\mathrm{T}}_{n-1,(\alpha)} - \mathrm{T}_{\mathrm{jack}})(\mathrm{T}_{\mathrm{jack}} - \mathrm{T}_n) + (\mathrm{T}_{\mathrm{jack}} - \mathrm{T}_n)^2 \right] \\
&= \frac{1}{n(n-1)} \left[\sum_{\alpha=1}^{n} (\tilde{\mathrm{T}}_{n-1,(\alpha)} - \mathrm{T}_{\mathrm{jack}})^2 + 2(\mathrm{T}_{\mathrm{jack}} - \mathrm{T}_n) \sum_{\alpha=1}^{n} (\tilde{\mathrm{T}}_{n-1,(\alpha)} - \mathrm{T}_{\mathrm{jack}}) \right. \\
&\quad \left. + \sum_{\alpha=1}^{n} (\mathrm{T}_{\mathrm{jack}} - \mathrm{T}_n)^2 \right] \\
&= v_1 + \frac{1}{n-1} (\mathrm{T}_{\mathrm{jack}} - \mathrm{T}_n)^2 + 2(\mathrm{T}_{\mathrm{jack}} - \mathrm{T}_n)(n\mathrm{T}_{\mathrm{jack}} - n\mathrm{T}_{\mathrm{jack}}) \\
&= v_1 + \frac{1}{n-1} (b_{\mathrm{jack}})^2.
\end{aligned}
$$

This shows that v_1 and v_2 are asymptotically the same as $n \to \infty$. It also shows that v_2 is not the same as $v_1 + (b_{\mathrm{jack}})^2$ and hence it is not an estimate of the mean square error of T_n, as is sometimes incorrectly claimed.

Applications To The Present Study

Since the DE, reweighted, and difference estimators are always based on the full sample, their bias is correctly estimated by b_{jack} . That is, if θ_k is the estimate based on all k of the jackknife replicate data sets and $\theta_{k-1,(\alpha)}$ is the estimate based on all but the αth replicated data set, then the bias of any one of these three estimators can be correctly estimated by

$$\mathrm{bias}(\theta_k) = (k-1)(\bar{\theta}_{k-1,(\cdot)} - \theta_k),$$

where in the present application $k = 15$ and $\bar{\theta}_{k-1,(\cdot)}$ is the mean of the $\theta_{k-1,(\alpha)}$ over k. However, the jackknife bias estimator is useless in estimating the bias of the list-based estimators, since they are not based on a sample of the complete population (relying instead only on the list sample or the list plus July NOL sample).

Since the direct expansion estimator D_k is unbiased with respect to sampling, the bias of any of the list-based estimators θ_k can be written as:

$$\mathrm{bias}(\theta_k) = \mathrm{E}(\theta_k) - \theta = \mathrm{E}(\theta_k) - \mathrm{E}(\mathrm{D}_k) = \mathrm{E}(\theta_k - \mathrm{D}_k).$$

Setting $\mathrm{T}_k = \theta_k - \mathrm{D}_k$, provides an estimate of the bias of θ_k , $\widehat{\mathrm{bias}}(\theta_k)$. The jackknife bias of the estimated bias, $\widehat{\mathrm{bias}}(\theta_k)$, is given by:

$$b_{\mathrm{jack}}(\widehat{\mathrm{bias}}(\theta_k)) = (k-1)\left[(\theta_k - \mathrm{D}_k) - (\bar{\theta}_{k-1,(\cdot)} - \bar{\mathrm{D}}_{k-1,(\cdot)}) \right].$$

The variance of the estimated bias of T_k, $\hat{\text{bias}}(\theta_k)$, can be estimated using either formula v_1 or v_2. The resulting estimates are given by:

$$\begin{aligned} v_1(\hat{\text{bias}}(\theta_k)) &= \frac{k-1}{k}\sum_{\alpha=1}^{k}(T_{k-1,(\alpha)} - \bar{T}_{k-1,(\cdot)})^2 \\ &= \frac{k-1}{k}\sum_{\alpha=1}^{k}\left[(\theta_{k-1,(\alpha)} - D_{k-1,(\alpha)}) - (\bar{\theta}_{k-1,(\cdot)} - \bar{D}_{k-1,(\,)})\right]^2. \end{aligned}$$

and

$$\begin{aligned} v_2(\hat{\text{bias}}(\theta_k)) &= \frac{k-1}{k}\sum_{\alpha=1}^{k}(T_{k-1,(\alpha)} - T_k)^2 \\ &= \frac{k-1}{k}\sum_{\alpha=1}^{k}\left[(\theta_{k-1,(\alpha)} - D_{k-1,(\alpha)}) - (\theta_k - D_k)\right]^2 \end{aligned}$$