United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research Division

SRB Research Report
Number SRB-93-04

July 1993

GENERALIZED POST-STRATIFIED ESTIMATORS IN THE AGRICULTURAL LABOR SURVEY

Charles R. Perry
Raj S. Chhikara
Lih-Yuan Deng
William C. Iwig
Scott Rumburg

GENERALIZED POST-STRATIFIED ESTIMATORS IN THE AGRICULTURAL LABOR SURVEY, By Charles R. Perry, Raj S. Chhikara*, Lih-Yuan Deng**, William C. Iwig and Scot Rumburg, Sampling and Estimation Research Section, Survey Research Branch, Research Division, National Agricultural Statistics Service, United States Department of Agriculture, Washington DC 20250-2000, July 1993, Report No. SRB-93-04.

ABSTRACT

The design-based characteristics of a general post-stratified estimator are investigated. Each of the four post-stratified estimators of current interest to NASS is a special case of the general post-stratified estimator. An approximation to the design-variance of the general estimator is derived using Taylor series methodology.

A simulation study is performed to evaluate the relative efficiency of certain list-only type post-stratified estimators. The approximate variance formula for the generalized post-stratified estimator is evaluated. The numerical evaluations show that the performance of a post-stratified estimator is largely a function of the sample size used to estimate the post-stratum weights, the sample size used to estimate the post-stratum means of the variable of interest, and the ratio of these two sample sizes. The relative efficiency increases as this ratio of two sample sizes increases. The approximate variance formula is found to be reasonably accurate for moderate size samples and highly accurate for large size samples.

KEYWORDS

Taylor series expansion; Design-variance; Post-stratified estimator; Simulation; Bias; Relative efficiency.

ACKNOWLEDGEMENTS

We thank Fred Vogel for putting forward the "strawman" concept and thus stimulating our thinking about new approaches to estimation. We also thank Ron Bosecker, Jim Davies and George Hanuschak for their helpful suggestions, Phil Kott for his technical review of this report and Jiannong Wang for his assistance in conducting the simulation study.

*Raj S. Chhikara is Professor, Division of Computing and Mathematics, University of Houston-Clear Lake.
**Lih-Yuan Deng is Associate Professor, Department of Mathematical Sciences, Memphis State University.

TABLE OF CONTENTS

SUMMARY

A post-stratification approach to estimation in follow-on surveys is formulated in the context of the Agricultural Labor Survey. A generalized post-stratified estimator is defined that makes use of the area and list frame samples from both the June Agricultural Survey and the follow-on Agricultural Labor Surveys. The approach includes as special cases two sets of post-stratified estimators of interest to NASS: (1) multiple-frame post-stratified estimators that use both area and list respondents from the follow-on survey and (2) list-only post-stratified estimators that use the list respondents from the follow-on survey exclusively. For each special case two possibilities for estimating the post-stratum means are included under the generalized approach: (1) the unweighted average response and (2) the weighted average response where the weights are the expansion factors associated with the sample units. Consideration of the special cases is motivated by the desire to use only list samples in follow-on survey.

The complexity of the multiple-frame sample design makes the variance of the generalized post-stratified estimator somewhat intractable. A computational formula for estimating the variance in the most general multiple-frame setting is derived using the Taylor series methodology. The approximate variance formula is extended by analogy to obtain estimators for the variance of the design-bias and the variance of the ratio of two general post-stratified estimators from two occasions.

An extensive simulation study is performed to evaluate numerically the performance of certain list-only post-stratified estimators. Both the bias and the relative efficiency of each estimator are evaluated. From the numerical results given in the paper, it follows that the post-stratification improves upon the precision of an estimator, provided that the sample size in a follow-on labor survey is moderate to large and that the variable used in post-stratifying the agricultural operations is reasonably correlated with the variable of interest. Also the performance of a post-stratified estimator is heavily dependent upon the sample size used in estimating post-stratum sizes. The relative efficiency of an estimator relative to the direct expansion estimator (computed using their mean square errors) increases as the size of the larger sample from a base period survey increases relative to

the size of the smaller sample from a follow-on survey. The numerical evaluations of the approximate variance formula show it to be fairly accurate for moderate sample sizes and highly accurate for large samples as one would expect.

INTRODUCTION

A multiple-frame sampling methodology is the basis of agricultural surveys conducted by the National Agricultural Statistics Service (NASS) of the U.S. Department of Agricultural (USDA). This methodology consists of sampling both a list frame and an area frame with the two samples drawn independently. List sampling is much more convenient and efficient compared to the area frame sampling. The former however does not provide a full coverage of all the agricultural operations. Thus, the area frame samples are used to compensate for the list undercoverage. The estimation method requires determination of the overlap or nonoverlap (OL/NOL) areas between the list and area samples. Since this OL/NOL delineation process is labor intensive and the NOL involves many fewer samples, its component of an estimate is often quite unreliable and is produced at an undesirable high cost.

In his "strawman" proposal, Vogel (1990) advocated a new approach to sampling and estimation for NASS surveys. One of the key ideas underlying this proposal is the use of post-stratified estimators in follow-on surveys. The use of post-stratified estimators is motivated by a desire to produce list-only estimates and a desire to eliminate the OL/NOL delineation. The desire to produce list-only estimates results primarily from the need to reduce respondent burden on NOL operations. The desire to eliminate the OL/NOL delineation results from the assumption that the overall quality of the survey would be improved by eliminating the numerous non-sampling errors associated with the complexity of the OL/NOL delineation. The word "strawman" in the proposal title implies that the post-stratification should be viewed as a baseline approach for developing improved estimation.

This idea has stimulated us to evaluate certain post-stratified estimators that might be useful in achieving the above objectives. The next section of this paper formulates a general post-stratified estimator in the context of the NASS Agricultural Labor Survey and explains how each of the post-stratified estimators of current interest to NASS is

obtained as a special case. A design-based computational formula for estimating the variance of the general post-stratified estimator is derived using Taylor series methodology. More importantly, this paper develops a set of rather general design-based variance and bias formulas for evaluating post-stratified estimators with respect to the sampling designs used by NASS.

A simulation study is performed to evaluate the relative efficiency of certain post-stratified estimators. The approximate variance formula derived from a Taylor series expansion is evaluated for its accuracy and hence, its appropriateness for estimating the variance of the general post-stratified estimator.

In Appendix A, we show how to write the obvious estimator of the design-bias of the general post-stratified estimator in a form analogous to the estimator itself. We then show how to use the general variance formula to estimate the variance of such an estimated design-bias.

A companion report (Rumburg, Perry, Chhikara and Iwig, 1993) provides detailed design-based empirical evaluations of the multiple frame and list-only post-stratified estimators of the number of hired workers for the NASS Agricultural Labor Survey using the 1991-92 Labor Survey and June Agricultural Survey data for the states of Florida and California. In this application, the post-stratum sizes were estimated using only all the area samples from the June Agricultural Survey.

A GENERALIZED POST-STRATIFIED ESTIMATOR

In the context of a follow-on agricultural survey, a generalized post-stratified estimator of a characteristic of interest, say population total in a state or region, is of the form:

$$\hat{G} = \sum_{k=1}^{K} \hat{G}_k, \tag{1}$$

where, for the post-stratum k,

$$\hat{G}_k = \frac{\hat{Z}_k \hat{Y}_k}{\hat{X}_k},$$

$\hat{Z}_k$ represents an estimate of the "size" of post-stratum k obtained from the base survey, e.g. June Agricultural Survey (JAS), $\hat{X}_k$ represents an estimate of the "size" of post-stratum k obtained from the follow-on survey, e.g. an Agricultural Labor Survey (ALS), and $\hat{Y}_k$ represents an estimate of the total, Y_k , for the item of interest for post-stratum k, derived from the follow-on survey.

For application to the ALS (i.e., follow-on labor surveys) there are four post-stratified estimators of primary interest as studied by Rumburg, et al. (1993). If we let $\hat{M}_k$ represent an estimate of the population size for post-stratum k derived from the JAS (the base survey), then by a simple transformation of the data the four estimators can be written as:

$$\hat{Y}_{G(List+NOL)wt} = \sum \hat{M}_k \bar{y}_{k(List+NOL)wt}, \tag{1.1}$$

where $\bar{y}_{k(List+NOL)wt}$ denotes the weighted mean of all List and NOL sample responses that are in post-stratum k,

$$\hat{Y}_{G(List+NOL)} = \sum \hat{M}_k \bar{y}_{k(List+NOL)}, \tag{1.2}$$

where $\bar{y}_{k(List+NOL)}$ denotes the simple mean of all List and NOL sample responses that are in post-stratum k,

$$\hat{Y}_{G(List)wt} = \sum \hat{M}_k \bar{y}_{k(List)wt}, \tag{1.3}$$

where $\bar{y}_{k(List)wt}$ denotes the weighted mean of all List sample responses that are in post-stratum k, and

$$\hat{Y}_{G(List)} = \sum \hat{M}_k \bar{y}_{k(List)}, \tag{1.4}$$

where $\bar{y}_{k(List)}$ denotes the simple mean of all List samples that are in post-stratum k.

Each of these estimators (1.1)-(1.4) is easily written in the notation of the general post-stratified estimator. For example, the estimator (1.1) is put in the notation of the general form, by equating $\hat{M}_k$ with $\hat{Z}_k$ and then equating $\bar{y}_{k(List+NOL)wt}$ with $\hat{Y}_k$ divided by $\hat{X}_k$, since in the latter case, it equals the sum of the expanded y's in post-stratum k divided by the sum of associated weights. Similarly follows the description of other estimators in the notation of the general estimator.

Cochran (1977, pages 142 -144) considered the problem of estimating totals and means for domains. If this estimation is extended to the whole population, one has the estimator as in (1.3). Särndal, Swensson and Wretman (1991, page 268) considered a estimator similar to (1.3) with the post-stratum size M_k assumed known. Hence, their estimator is a special case of the generalized post-stratified estimator defined in Equation (1).

TAYLOR SERIES VARIANCE APPROXIMATION

Following the standard Taylor linearization method (Wolter 1985, page 226), the first-order approximation of the design-variance of the generalized post-stratified estimator $\hat{G}$ given in Equation (1) can be written as:

$$V(\hat{G}) = V\left(\sum_{k=1}^{K} \partial G_{Z_k}\hat{Z}_k + \partial G_{Y_k}\hat{Y}_k + \partial G_{X_k}\hat{X}_k\right), \tag{2}$$

where

$$\partial G_{Z_k} = \frac{\partial G}{\partial Z_k}\Big|_{(\vec{Z},\vec{Y},\vec{X})} = \frac{Y_k}{X_k},$$
$$\partial G_{Y_k} = \frac{\partial G}{\partial Y_k}\Big|_{(\vec{Z},\vec{Y},\vec{X})} = \frac{Z_k}{X_k},$$
$$\partial G_{X_k} = \frac{\partial G}{\partial X_k}\Big|_{(\vec{Z},\vec{Y},\vec{X})} = -\frac{Z_k Y_k}{X_k^2}.$$

and

$$\vec{Z} = (Z_1, \cdots, Z_K) = (E(\hat{Z}_1), \cdots, E(\hat{Z}_K)),$$
$$\vec{Y} = (Y_1, \cdots, Y_K) = (E(\hat{Y}_1), \cdots, E(\hat{Y}_K)),$$
$$\vec{X} = (X_1, \cdots, X_K) = (E(\hat{X}_1), \cdots, E(\hat{X}_K)).$$

In the general form of the post-stratified estimator, each $\hat{Z}_k, \hat{Y}_k$ and $\hat{X}_k$ will have a list and an area frame component. Hence these estimators can be written as:

$$\begin{aligned} \hat{Z}_k &= \hat{Z}_{Lk} + \hat{Z}_{Ak}, \\ \hat{Y}_k &= \hat{Y}_{Lk} + \hat{Y}_{Ak}, \\ \hat{X}_k &= \hat{X}_{Lk} + \hat{X}_{Ak}, \end{aligned} \tag{3}$$

where the list and area frame components are denoted by subscripts L and A, respectively.

Since samples are drawn independently from the list and area frames, the list estimates $\hat{Z}_{Lk}, \hat{Y}_{Lk}$ and $\hat{X}_{Lk}$ are independent of the area estimates $\hat{Z}_{Ak}, \hat{Y}_{Ak}$ and $\hat{X}_{Ak}$. Thus the Taylor series variance approximation can be rewritten as:

$$\begin{aligned} V(\hat{G}) = V &\left(\sum_{k=1}^{K} \partial G_{Z_k} \hat{Z}_{Lk} + \partial G_{Y_k} \hat{Y}_{Lk} + \partial G_{X_k} \hat{X}_{Lk} \right) \\ + V &\left(\sum_{k=1}^{K} \partial G_{Z_k} \hat{Z}_{Ak} + \partial G_{Y_k} \hat{Y}_{Ak} + \partial G_{X_k} \hat{X}_{Ak} \right). \end{aligned} \tag{4}$$

Since the list frame sample for the Agricultural Labor Survey (i.e., for the follow-on current period survey) is drawn independently of the list frame sample for the June Agricultural Survey (i.e., for the base survey), the list estimate $\hat{Z}_{Lk}$ is independent of the other list estimates $\hat{Y}_{Lk}$ and $\hat{X}_{Lk}$. Thus the first term of Equation (4) can be rewritten as:

$$\begin{aligned} V &\left(\sum_{k=1}^{K} \partial G_{Z_k} \hat{Z}_{Lk} + \partial G_{Y_k} \hat{Y}_{Lk} + \partial G_{X_k} \hat{X}_{Lk} \right) \\ &= V \left(\sum_{k=1}^{K} \partial G_{Z_k} \hat{Z}_{Ljk} \right) + V \left(\sum_{k=1}^{K} \partial G_{Y_k} \hat{Y}_{L\mathcal{L}k} + \partial G_{X_k} \hat{X}_{L\mathcal{L}k} \right), \end{aligned} \tag{5}$$

where on the right side of the equation the subscript L_J represents the June Agricultural Survey list and the subscript $L_{\mathcal{L}}$ represents the Labor Survey list.

The area frame component of the Labor Survey is based on approximately 40% of the June area frame sample replicates -- the so called, 60/40 sample split. All the replicates associated with rotational years 4 and 5 and occasionally a small fraction of the replicates associated with rotational years 1, 2, and 3 constitute the Labor Survey area frame. Since each June area frame replicate is drawn independently, except for the restriction imposed by sampling without replacement, $\hat{Y}_{Ak}$, $\hat{X}_{Ak}$ and the Labor Survey area frame component of $\hat{Z}_{Ak}$ are essentially independent of the replicates of the area frame component of $\hat{Z}_{Ak}$ that are not used in the Labor Survey. In fact, let

$$\hat{Z}_{Ahk} = \hat{Z}_{A_{J\cap\mathcal{L}}hk} + \hat{Z}_{A_{J\setminus\mathcal{L}}hk},$$

where the two components of $\hat{Z}_{Ak}$ for the original sampling stratum h are denoted respectively by $\hat{Z}_{A_{J\cap\mathcal{L}}hk}$ and $\hat{Z}_{A_{J\setminus\mathcal{L}}hk}$, $h = 1, 2, \ldots, H$. Then it is easy to see that:

$$V\left(\sum_{k=1}^{K}\hat{Z}_{Ahk}\right) = \left(\frac{1-f_{Ah}}{1-f_{A_{J\cap\mathcal{L}}h}}\right)V\left(\sum_{k=1}^{K}\hat{Z}_{A_{J\cap\mathcal{L}}hk}\right) + \left(\frac{1-f_{Ah}}{1-f_{A_{J\setminus\mathcal{L}}h}}\right)V\left(\sum_{k=1}^{K}\hat{Z}_{A_{J\setminus\mathcal{L}}hk}\right)$$

for original stratum h. Here f represents the sampling fraction. This formulation allows the samples to be drawn using a stratified random sampling design as is the case with most agricultural surveys.

Thus the second term of Equation (4) can be rewritten as:

$$\begin{aligned} &V\left(\sum_{k=1}^{K}\partial G_{Z_k}\hat{Z}_{Ak} + \partial G_{Y_k}\hat{Y}_{Ak} + \partial G_{X_k}\hat{X}_{Ak}\right) \\ &\quad = V\left(\sum_{k=1}^{K}\partial G_{Z_k}\hat{Z}_{A_{J\setminus\mathcal{L}}k}\right) + V\left(\sum_{k=1}^{K}\partial G_{Z_k}\hat{Z}_{A_{J\cap\mathcal{L}}k} + \partial G_{Y_k}\hat{Y}_{A_{\mathcal{L}}k} + \partial G_{X_k}\hat{X}_{A_{\mathcal{L}}k}\right), \quad (6) \end{aligned}$$

where on the right side of the equation the subscript $A_{J\cap\mathcal{L}}$ indicates aggregation of the expanded June area frame data over the replicates used in the Labor Survey, the subscript

$A_{J\setminus\mathcal{L}}$ indicates aggregation of the expanded June area frame data over the replicates not used in the Labor Survey and the subscript $A_{\mathcal{L}}$ indicates Labor Survey area frame expansion and aggregation. This equality assumes that appropriate adjustments are made to all variance calculations as discussed above to account for differences in finite population correction factors. Operationally these differences can be ignored since almost all area frame sampling fractions are less than one percent.

Replacing the right hand side of Equation (4) with the expression given in Equations (5) and (6), the Taylor series expansion variance of $\hat{G}$ becomes:

$$V(\hat{G}) = V\left(\sum_{k=1}^{K} \partial G_{Z_k} \hat{Z}_{L_J k}\right) + V\left(\sum_{k=1}^{K} \partial G_{Y_k} \hat{Y}_{L_{\mathcal{L}} k} + \partial G_{X_k} \hat{X}_{L_{\mathcal{L}} k}\right)$$
$$+ V\left(\sum_{k=1}^{K} \partial G_{Z_k} \hat{Z}_{A_{J\setminus\mathcal{L}} k}\right) + V\left(\sum_{k=1}^{K} \partial G_{Z_k} \hat{Z}_{A_{J\cap\mathcal{L}} k} + \partial G_{Y_k} \hat{Y}_{A_{\mathcal{L}} k} + \partial G_{X_k} \hat{X}_{A_{\mathcal{L}} k}\right), \tag{7}$$

where the subscript

(1) L_J represents data expansion and aggregation for the June Agricultural Survey list,

(2) $L_{\mathcal{L}}$ represents data expansion and aggregation for the Labor Survey list,

(3) $A_{J\cap\mathcal{L}}$ represents data expansion for the June Agricultural Survey area frame and aggregated over the replicates used in the Labor Surveys (replicates associated with rotational years 4 and 5 and a small fraction of the replicates associated with rotational years 1, 2, and 3),

(4) $A_{J\setminus\mathcal{L}}$ represents data expansion for the June Agricultural Survey area frame and aggregated over the replicates not used in Labor Surveys, and

(5) $A_{\mathcal{L}}$ represents Labor Survey area frame data expansion and aggregation.

To complete the derivation of the computational form of the Taylor series variance estimation formula, some additional notation is needed. For the list component of the

June Agricultural Survey, let

$$\hat{Z}_{L_J k} = \sum_{h \in H_{L_J}} \hat{Z}_{L_J hk},$$
$$\hat{Z}_{L_J hk} = W_{L_J h} \sum_{i \in n_{L_J h}} z_{L_J hi}\, \delta(k, L_J hi), \tag{8}$$

$$\delta(k, L_J hi) = \begin{cases} 1, & \text{if unit } i \text{ of JAS list stratum } h \text{ is in post-stratum } k \\ 0, & \text{otherwise.} \end{cases}$$

Note that the notation $\sum_{h \in H_S}$ denotes the summation over all strata corresponding to survey type S. For example, $S = L_J$ is the June survey using the list frame.

In Equation (8), the expansion factor $W_{L_J h} = N_{L_J h}/n_{L_J h}$, where $N_{L_J h}$ and $n_{L_J h}$ are respectively the population size and sample size for JAS list frame stratum h.

For the list component of the Labor Survey, let

$$\hat{Y}_{L_{\mathcal{L}} k} = \sum_{h \in H_{L_{\mathcal{L}}}} \hat{Y}_{L_{\mathcal{L}} hk},$$
$$\hat{Y}_{L_{\mathcal{L}} hk} = W_{L_{\mathcal{L}} h} \sum_{i \in n_{L_{\mathcal{L}} h}} \hat{y}_{L_{\mathcal{L}} hi}\, \delta(k, L_{\mathcal{L}} hi), \tag{9}$$

and

$$\hat{X}_{L_{\mathcal{L}} k} = \sum_{h \in H_{L_{\mathcal{L}}}} \hat{X}_{L_{\mathcal{L}} hk},$$
$$\hat{X}_{L_{\mathcal{L}} hk} = W_{L_{\mathcal{L}} h} \sum_{i \in n_{L_{\mathcal{L}} h}} \hat{x}_{L_{\mathcal{L}} hi}\, \delta(k, L_{\mathcal{L}} hi), \tag{10}$$

where

$$\delta(k, L_{\mathcal{L}} hi) = \begin{cases} 1, & \text{if unit } i \text{ of Labor list stratum } h \text{ is in post-stratum } k \\ 0, & \text{otherwise.} \end{cases}$$

In Equations (9) and (10), the expansion factor $W_{L_{\mathcal{L}} h} = N_{L_{\mathcal{L}} h}/n_{L_{\mathcal{L}} h}$, where $N_{L_{\mathcal{L}} h}$ and $n_{L_{\mathcal{L}} h}$ are respectively the population size and sample size for Labor list frame stratum h.

For the area frame component of the June Agricultural Survey, let

$$\begin{aligned}
\hat{Z}_{A_{J\cap\mathcal{L}}k} &= \sum_{h\in H_{A_{J\cap\mathcal{L}}}} \hat{Z}_{A_{J\cap\mathcal{L}}hk}, \\
\hat{Z}_{A_{J\cap\mathcal{L}}hk} &= W_{A_Jh} \sum_{i\in n_{A_{J\cap\mathcal{L}}h}} z_{A_{J\cap\mathcal{L}}hik}, \\
z_{A_{J\cap\mathcal{L}}hik} &= \sum_{j\in M_{A_{J\cap\mathcal{L}}hi}} z_{A_{J\cap\mathcal{L}}hij}\,\delta(k, A_{J\cap\mathcal{L}}hij),
\end{aligned} \tag{11}$$

where

(1) $\sum_{i\in n_{A_{J\cap\mathcal{L}}h}}$ indicates the sum is over all $n_{A_{J\cap\mathcal{L}}h}$ June area frame sample segments of stratum h that are used in the Labor Survey area frame,

(2) $\sum_{j\in M_{A_{J\cap\mathcal{L}}hi}}$ indicates the sum is over all $M_{A_{J\cap\mathcal{L}}hi}$ tracts of sample segment $A_{J\cap\mathcal{L}}hi$,

(3)

$$\delta(k, A_{J\cap\mathcal{L}}hij) = \begin{cases} 1, & \text{if tract } j \text{ of segment } A_{J\cap\mathcal{L}}hi \text{ is in post-stratum } k \\ 0, & \text{otherwise} \end{cases}$$

and

$$\begin{aligned}
\hat{Z}_{A_{J\setminus\mathcal{L}}k} &= \sum_{h\in H_{A_{J\setminus\mathcal{L}}}} \hat{Z}_{A_{J\setminus\mathcal{L}}hk}, \\
\hat{Z}_{A_{J\setminus\mathcal{L}}hk} &= W_{A_Jh} \sum_{i\in n_{A_{J\setminus\mathcal{L}}h}} z_{A_{J\setminus\mathcal{L}}hik}, \\
z_{A_{J\setminus\mathcal{L}}hik} &= \sum_{j\in M_{A_{J\setminus\mathcal{L}}hi}} z_{A_{J\setminus\mathcal{L}}hij}\,\delta(k, A_{J\setminus\mathcal{L}}hij),
\end{aligned} \tag{12}$$

where

(1) $\sum_{i\in n_{A_{J\setminus\mathcal{L}}h}}$ indicates the sum is over all $n_{A_{J\setminus\mathcal{L}}h}$ June area frame sample segments of stratum h that are not used in the Labor Survey area frame,

(2) $\sum_{j\in M_{A_{J\setminus\mathcal{L}}hi}}$ indicates the sum is over all $M_{A_{J\setminus\mathcal{L}}hi}$ tracts of segment $A_{J\setminus\mathcal{L}}hi$,

(3)

$$\delta(k, A_{J\setminus\mathcal{L}}hij) = \begin{cases} 1, & \text{if tract } j \text{ of segment } M_{A_{J\setminus\mathcal{L}}hi} \text{ is in post-stratum } k \\ 0, & \text{otherwise.} \end{cases}$$

In Equations (11) and (12), the expansion factor $W_{AJh} = N_{AJh}/n_{AJh}$, where N_{AJh} and n_{AJh} are respectively the population size and sample size for June area frame stratum h.

For the area frame component of the Labor Survey, let

$$\begin{aligned} \hat{Y}_{A_{\mathcal{L}}k} &= \sum_{h\in H_{A_{\mathcal{L}}}} \hat{Y}_{A_{\mathcal{L}}hk}, \\ \hat{Y}_{A_{\mathcal{L}}hk} &= W_{A_{\mathcal{L}}h} \sum_{i\in n_{A_{\mathcal{L}}h}} \hat{y}_{A_{\mathcal{L}}hik}, \\ \hat{y}_{A_{\mathcal{L}}hik} &= \sum_{j\in m_{A_{\mathcal{L}}hi}} W_{sA_{\mathcal{L}}hij} y_{A_{\mathcal{L}}hij}\, \delta(k, A_{\mathcal{L}}hij) \end{aligned} \tag{13}$$

and

$$\begin{aligned} \hat{X}_{A_{\mathcal{L}}k} &= \sum_{h\in H_{A_{\mathcal{L}}}} \hat{X}_{A_{\mathcal{L}}hk}, \\ \hat{X}_{A_{\mathcal{L}}hk} &= W_{A_{\mathcal{L}}h} \sum_{i\in n_{A_{\mathcal{L}}h}} \hat{x}_{A_{\mathcal{L}}hik}, \\ \hat{x}_{A_{\mathcal{L}}hik} &= \sum_{j\in m_{A_{\mathcal{L}}hi}} W_{sA_{\mathcal{L}}hij} x_{A_{\mathcal{L}}hij}\, \delta(k, A_{\mathcal{L}}hij), \end{aligned} \tag{14}$$

where

(1) $\sum_{i\in n_{A_{\mathcal{L}}h}}$ indicates the sum is over all $n_{A_{\mathcal{L}}h}$ Labor area frame segments of stratum h (June stratum h segments used in Labor Survey),

(2) $\sum_{j\in m_{A_{\mathcal{L}}hi}}$ indicates the sum is over all $m_{A_{\mathcal{L}}hi}$ NOL tracts of segment $A_{\mathcal{L}}hi$,

(3)

$$\delta(k, A_{\mathcal{L}}hij) = \begin{cases} 1, & \text{if NOL tract } j \text{ of segment } A_{\mathcal{L}}hi \text{ is in post-stratum } k \\ 0, & \text{otherwise.} \end{cases}$$

In Equations (13) and (14), the first phase expansion factor $W_{A_{\mathcal{L}}h} = N_{A_{\mathcal{L}}h}/n_{A_{\mathcal{L}}h} = N_{A_{\mathcal{J}}h}/n_{A_{\mathcal{L}}h}$, where $N_{A_{\mathcal{L}}h} = N_{A_{\mathcal{J}}h}$ and $n_{A_{\mathcal{L}}h}$ are respectively the population size and sample size for Labor area frame stratum h. The second phase expansion factor for tract j from segment i and stratum h which is selected in second phase sample from post-select stratum s is $W_{sA_{\mathcal{L}}hij} = N_{sA_{\mathcal{L}}h}/n_{sA_{\mathcal{L}}h}$, where $N_{sA_{\mathcal{L}}h}$ and $n_{sA_{\mathcal{L}}h}$ are respectively the population size and sample size for post-select stratum s.

By substituting Equations (8) through (14) in Equation (7) and rearranging the order of summation the first-order Taylor series variance can be rewritten as:

$$
\begin{aligned}
V(\hat{G}) = {} & V_{L_{\mathcal{J}}}\left(\sum_{h\in H_{L_{\mathcal{J}}}} W_{L_{\mathcal{J}}h} \sum_{i\in n_{L_{\mathcal{J}}h}} \left(\sum_{k=1}^{K} \partial G_{Z_k} z_{L_{\mathcal{J}}hi}\, \delta(k, L_{\mathcal{J}}hi)\right)\right) + \\
& V_{L_{\mathcal{L}}}\left(\sum_{h\in H_{L_{\mathcal{L}}}} W_{L_{\mathcal{L}}h} \sum_{i\in n_{L_{\mathcal{L}}h}} \left(\sum_{k=1}^{K} [\partial G_{Y_k} y_{L_{\mathcal{L}}hi} + \partial G_{X_k} x_{L_{\mathcal{L}}hi}]\, \delta(k, L_{\mathcal{L}}hi)\right)\right) + \\
& V_{A_{\mathcal{J}\setminus\mathcal{L}}}\left(\sum_{h\in H_{A_{\mathcal{J}\setminus\mathcal{L}}}} W_{A_{\mathcal{J}}h} \sum_{i\in n_{A_{\mathcal{J}\setminus\mathcal{L}}h}} \sum_{j\in M_{A_{\mathcal{J}\setminus\mathcal{L}}hi}} \left(\sum_{k=1}^{K} \partial G_{Z_k} z_{A_{\mathcal{J}\setminus\mathcal{L}}hij}\, \delta(k, A_{\mathcal{J}\setminus\mathcal{L}}hij)\right)\right) + \\
& V_{A_{\mathcal{L}}}\left(\sum_{h\in H_{A_{\mathcal{J}\cap\mathcal{L}}}} W_{A_{\mathcal{J}}h} \left[\sum_{i\in n_{A_{\mathcal{J}\cap\mathcal{L}}h}} \sum_{j\in M_{A_{\mathcal{J}\cap\mathcal{L}}hi}} \left(\sum_{k=1}^{K} \partial G_{Z_k} z_{A_{\mathcal{J}\cap\mathcal{L}}hij}\, \delta(k, A_{\mathcal{J}\cap\mathcal{L}}hij)\right)\right] + \right. \\
& \left. \sum_{h\in H_{A_{\mathcal{L}}}} W_{A_{\mathcal{J}}h} \sum_{i\in n_{A_{\mathcal{L}}h}} \left[\sum_{j\in m_{A_{\mathcal{L}}hi}} W_{sA_{\mathcal{L}}hij} \left(\sum_{k=1}^{K} [\partial G_{Y_k} y_{A_{\mathcal{L}}hij} + \partial G_{X_k} x_{A_{\mathcal{L}}hij}]\, \delta(k, A_{\mathcal{L}}hij)\right)\right]\right).
\end{aligned}
\tag{15}
$$

The sequence of computations necessary to produce the first- order Taylor series variance estimate are indicated more clearly by rewriting Equation (15) as:

$$
\begin{aligned}
V(\hat{G}) = {} & \sum_{h\in H_{L_{\mathcal{J}}}} V_{L_{\mathcal{J}}}\left(\sum_{i\in n_{L_{\mathcal{J}}h}} t_{L_{\mathcal{J}}hi}\right) + \sum_{h\in H_{L_{\mathcal{L}}}} V_{L_{\mathcal{L}}}\left(\sum_{i\in n_{L_{\mathcal{L}}h}} t_{L_{\mathcal{L}}hi}\right) + \\
& \sum_{h\in H_{A_{\mathcal{J}\setminus\mathcal{L}}}} V_{A_{\mathcal{J}\setminus\mathcal{L}}}\left(\sum_{i\in n_{A_{\mathcal{J}\setminus\mathcal{L}}h}} t_{A_{\mathcal{J}\setminus\mathcal{L}}hi}\right) + \sum_{h\in H_{A_{\mathcal{L}}}} V_{A_{\mathcal{L}}}\left(\sum_{i\in n_{A_{\mathcal{L}}h}} (t_{A_{\mathcal{J}\cap\mathcal{L}}hi} + \hat{t}_{A_{\mathcal{L}}hi})\right),
\end{aligned}
\tag{16}
$$

where

$$t_{L_J hi} = W_{L_J h} \sum_{k=1}^{K} \partial G_{Z_k} z_{L_J hi}\, \delta(k, L_J hi),$$

$$t_{L_{\mathcal{L}} hi} = W_{L_{\mathcal{L}} h} \sum_{k=1}^{K} [\partial G_{Y_k} y_{L_{\mathcal{L}} hi} + \partial G_{X_k} x_{L_{\mathcal{L}} hi}]\, \delta(k, L_{\mathcal{L}} hi),$$

$$t_{A_{J\setminus\mathcal{L}} hi} = W_{A_J h} \sum_{j \in M_{A_{J\setminus\mathcal{L}} hi}} \sum_{k=1}^{K} \partial G_{Z_k} z_{A_{J\setminus\mathcal{L}} hij}\, \delta(k, A_{J\setminus\mathcal{L}} hij),$$

$$t_{A_{J\cap\mathcal{L}} hi} = W_{A_J h} \sum_{j \in M_{A_{J\cap\mathcal{L}} hi}} \sum_{k=1}^{K} \partial G_{Z_k} z_{A_{J\cap\mathcal{L}} hij}\, \delta(k, A_{J\cap\mathcal{L}} hij),$$

$$\hat{t}_{A_{\mathcal{L}} hi} = W_{A_{\mathcal{L}} h} \sum_{j \in m_{A_{\mathcal{L}} hi}} W_{sA_{\mathcal{L}} hij} \sum_{k=1}^{K} \left[\partial G_{Y_k} y_{A_{\mathcal{L}} hij} + \partial G_{X_k} x_{A_{\mathcal{L}} hij}\right]\, \delta(k, A_{\mathcal{L}} hij).$$

The general form of the formula for computing the stratum level variance estimates in Equation (16) is given by:

$$V_h \left(\sum_{i \in n_h} t_{hi} \right) = \left(\frac{N_h - n_h}{N_h} \right) \left(\frac{n_h}{n_h - 1} \right) \left(\sum_{i \in n_h} t_{hi}^2 - \frac{\left(\sum_{i \in n_h} t_{hi} \right)^2}{n_h} \right), \qquad (17)$$

where the stratum population and sample sizes are determined by the subscripts as these appear in Equations (16) and as previously defined in Equations (7)-(14).

The use of Equation (17) to compute the stratum level variances is discussed in Kott (1990b). Equation (17) provides unbiased stratum level variance estimates for the first and second terms of Equation (16) and for the third term when the finite population correction factor is adjusted as indicated in the discussion preceding Equation (6). Since the fourth term of Equation (16) involves a second phase of sampling, the stratum level estimates derived from Equation (17) will be slightly conservative, [see Kott (1990b) pages 19-22, particularly Equation (26) on page 22, therein]. Operationally, the finite population correction factors in Equation (17) are generally ignored. The effect of ignoring the finite population correction factors on the variance estimates is not important since the sampling rate in most strata is very small.

EMPIRICAL SIMULATION EVALUATIONS

In this section, we evaluate the performance of the generalized post-stratified estimator and its variance formula discussed in the previous section by conducting a simulation study. The NASS sampling design involves separate stratifications for the list and area frames. So it is desirable to simulate data for the two frames separately for the evaluation. In the post-stratification approach described above, the JAS serves as a base survey and the ALS as a follow-on survey. As such we have used this base and follow-on survey format to simulate the data. Although we consider simulations based only on a single type of sampling frame and hence, do not exactly simulate the two separate stratifications analogous to NASS, it however emulates the basic approach adopted in the construction of the generalized post-stratified estimator which makes use of JAS sample responses to estimate post-stratum sizes and ALS sample responses to estimate the post-stratum means. The sampling procedure and estimators considered for simulation evaluations are described next.

Description of sampling procedure for simulation.

The sampling procedure for simulation is based on the concept that the original stratification is common to both the base survey and the follow-on survey. This approach is general enough to be applicable to many situations involving follow-on surveys. Presently we consider the stratification for the list frame and evaluate the post-stratification methodology for the list-only estimators. Therefore, the subscript L (for list frame) will be dropped from the notations in previous sections.

(1) Suppose there is a sample of size n_{Jh}, representing the JAS sample values, from the list frame which consists of H original strata with N_{Jh} as the size of h-th stratum, $h = 1, 2, ..., H$.

(2) After a JAS sample is taken, observations are post-stratified into K post-strata according to some stratification variables with some observed charac-

teristics. Population counts for the post-strata are produced once annually from the JAS and then these counts are fixed for the ALS.

(3) $m_{Jh,k}$ is the number of units in the h-th original stratum classified into the k-th post-stratum in the JAS sample.

(4) An independent sample of size $n_{\mathcal{L}h}$, representing the ALS sample values, is drawn from the list frame.

(5) After an independent ALS sample is taken, we post-stratify observations, according to the stratification variable, into K post-strata same as those for the JAS sample.

(6) $m_{\mathcal{L}h,k}$ is the number of units in the h-th original stratum classified into the k-th post-stratum in the ALS sample.

Estimators under consideration for simulation.

We denote by $\hat{Y}_G$ the generalized post-stratified estimator given in Equation (1), that is

$$\hat{Y}_G = \sum_{k=1}^{K} \hat{G}_k, \tag{18}$$

where

$$\hat{G}_k = \frac{\hat{Z}_k \hat{Y}_k}{\hat{X}_k},$$

for post-stratum k. In terms of the above notations, $\hat{Z}_k$, $\hat{X}_k$ and $\hat{Y}_k$ can be expressed as:

$$\hat{Z}_k = \sum_{h=1}^{H} \frac{N_{Jh}}{n_{Jh}} m_{Jh,k} \tag{19}$$

an estimate of the "size" of post-stratum k derived from JAS,

$$\hat{X}_k = \sum_{h=1}^{H} \frac{N_{Jh}}{n_{\mathcal{L}h}} m_{\mathcal{L}h,k} \tag{20}$$

represents an estimate of the "size" of post-stratum k derived from ALS, and

$$\hat{Y}_k = \sum_{h=1}^{H} \frac{N_{Jh}}{n_{\mathcal{L}h}} \sum_{i \in S_k} y_{hi} \tag{21}$$

represents an estimate of the total, Y_k , for the item of interest for post-stratum k, derived from the ALS.

Note that the generalized post-stratified estimator can also be re-written as

$$\hat{Y}_{G\,wt} = \sum_{k=1}^{K} \hat{M}_k \hat{\bar{Y}}_k, \tag{22}$$

where

$$\hat{M}_k = \hat{Z}_k = \sum_{h=1}^{H} \frac{N_{Jh}}{n_{Jh}} m_{Jh,k}$$

and

$$\hat{\bar{Y}}_k = \frac{\sum_{h=1}^{H} \frac{N_{Jh}}{n_{\mathcal{L}h}} \sum_{i \in S_k} y_{hi}}{\sum_{h=1}^{H} \frac{N_{Jh}}{n_{\mathcal{L}h}} m_{\mathcal{L}h,k}} = \frac{\hat{Y}_k}{\hat{X}_k}.$$

Since $\hat{\bar{Y}}_k$ is a weighted mean of the sample observations in post-stratum k, $\hat{Y}_{G\,wt}$ corresponds to the weighted post-stratified estimator given in Equation (1.3).

The other post-stratified estimator considered is one that uses the simple mean of the sample observations in a post-stratum:

$$\hat{Y}_{G\,unwt} = \sum_{k=1}^{K} \hat{M}_k \hat{\bar{Y}}_k, \tag{23}$$

where

$$\hat{M}_k = \sum_{h=1}^{H} \frac{N_{Jh}}{n_{Jh}} m_{Jh,k}, \quad \hat{\bar{Y}}_k = \bar{y}_k.$$

This corresponds to the unweighted post-stratified estimator given in Equation (1.4).

In this simulation study, another post-stratified estimator considered is based on the use of the combined expansion factor, i.e. the unweighted post-stratum count estimates, in addition to the unweighted post-stratum mean estimates. This unweighted "combined" estimator is:

$$\hat{Y}_{G\,unwt(C)} = \sum_{k=1}^{K} \hat{M}_k \hat{\bar{Y}}_k, \tag{24}$$

where

$$\hat{M}_k = \frac{\sum_{h=1}^{H} N_{Jh}}{\sum_{h=1}^{H} n_{Jh}} \sum_{h=1}^{H} m_{Jh,k}, \quad \hat{Y}_k = \bar{y}_k.$$

This estimator is motivated by the observation that there can be large variation in estimating the post-stratum count separately in each original stratum and so a combined population count estimate across all post-strata will stabilize this variation. The term "combined" is used here in the manner similar to the "combined ratio estimator" discussed in survey sampling literature. [See, for example, Cochran (1977, pages 164-169).]

For the sake of comparison, also considered is the direct expansion estimator of Y given by

$$\hat{Y}_{st} = \sum_{h=1}^{H} N_{Jh} \bar{y}_h. \tag{25}$$

Model for data generation.

The following steps are followed to generate a response variable y and a correlated auxiliary variable x to form a population of H original strata with stratum size N_{Jh} for $h = 1, 2, ..., H$, and another auxiliary variable w for post-stratification of population units. In the context of the ALS, the variable y represents the number of hired workers, x is the size of farm operation and w is the peak number of workers expected during the year for the farm labor.

(1) For each $h = 1, 2, ..., H$, $i = 1, 2, ..., N_{Jh}$, stratum data are generated as follows:

(a) First we generate a population of base values z_{hi} uniformly distributed over an interval, say

$$z_{hi} \sim U(Z_1, Z_2),$$

where U stands for the uniform distribution and interval (Z_1, Z_2) is chosen to be $(5, 8)$.

(b) The values for the auxiliary variable x representing the farm operation size are generated using the model,

$$x_{hi} = v_h + \alpha_h z_{hi} + \epsilon_2, \tag{26}$$

where $\epsilon_2 \sim N(0, \sigma_2^2)$, $v_h \sim U(V_1, V_2)$ and $\alpha_h \sim U(\Delta_1, \Delta_2)$. The parameter values chosen are: $\sigma_2^2 = 1$, $(V_1, V_2) = (3, 5)$ and $(\Delta_1, \Delta_2) = (0.5, 4)$.

Steps (a)-(b) would generate values that may vary due to size, stratum or other characteristics of population units. More specifically, z_{hi} represents the unit size, v_h represents the stratum mean, and α_h represents the dependence of the stratification variable on the unit size.

Next, we consider another auxiliary variable w to be used for post-stratification. This variable may be similar to x, but invariably it is expected to reflect additional information.

(c) Generate

$$x'_{hi} = z_{hi} + \epsilon_1, \quad \text{where} \quad \epsilon_1 \sim N(0, \sigma_1^2),$$

where σ_1^2 is chosen to be 1.

(d) From values generated in steps (b) and (c), generate

$$w_{hi} = p_h x'_{hi} + (1 - p_h) x_{hi}, \tag{27}$$

where p_h is generated randomly from the uniform distribution over an interval. Presently the interval is taken to be $(0.0, 0.1)$. This makes w_{hi} not to be too different from x_{hi}. This consideration is quite appropriate in the context of the ALS since the use of the peak number of workers as a stratification variable is reflected in the list frame stratification.

(e) For the response variable y, generate

$$y_{hi} = \mu_h + \beta_h w_{hi} + \epsilon_3, \tag{28}$$

where $\epsilon_3 \sim N(0, \sigma_3^2)$ and μ_h are selected randomly from $U(M_1, M_2)$ and β_h are selected randomly from $U(B_1, B_2)$. This model takes into account the differences in response due to stratum and other characteristics of population units.

(2) The variate w_{hi}, representing an auxiliary variable for post-stratification, is, in theory, a better stratification variable than the original stratification variable in order for the post-stratification to be efficient.

(3) Once w_{hi} are generated, we sort the values of w_{hi} and find the cut-off points for the post-stratification. (Say, $A_1 < A_2 < ... < A_{K-1}$ are the cut-off points so that if $w_{hi} \in (A_{k-1}, A_k)$, then the unit is classified as belonging to post-stratum k.)

Simulation parameters.

Several parameters (which, we think, will not affect the outcome much) are fixed in this simulation study:

(1) Population sizes are randomly generated from $U(500, 5000)$.

(2) The number of original strata $H = 6$ and the number of post strata , $K = 8$.

The total sample size for ALS is $n_{\mathcal{L}} = 50, 100, 200, 400$, where

$$n_{\mathcal{L}} = \sum_{h=1}^{H} n_{\mathcal{L}h}. \tag{29}$$

The relative size of the expansion factor $N_{Jh}/n_{\mathcal{L}h}$ is randomly generated; where we first generate randomly E_h from $U(50, 200)$, and then rescale E_h such that $\sum_{h=1}^{H} n_{\mathcal{L}h} = n_{\mathcal{L}}$. It is straightforward to compute the appropriate scaling constant c for the expansion

factors, $N_{Jh}/n_{\mathcal{L}h} = c * E_h$ for $h = 1, 2, \ldots, H$, so that

$$c = \frac{\sum_{h=1}^{H} N_{Jh}/E_h}{n}.$$

We then selected the following cases of the various parameters in Equation (28).

(1) The standard deviation of error term ϵ_3 is chosen to be $\sigma_3 = 1$ and $\sigma_3 = 2$.

(2) The intercept of μ_h is selected from the range $(M_1, M_2) = (5.0, 8.0)$.

(3) The slope of β_h is selected from the interval $(B_1, B_2) = (1.0, 2.0)$ and the interval $(B_1, B_2) = (3.0, 4.0)$.

In general, the post-stratification will be more effective than the original stratification, if (1) ϵ_3 is smaller, (2) the range of (M_1, M_2) is narrower and (3) the values of (B_1, B_2) are larger. However, when the range (M_1, M_2) was considered to be $(5.0, 6.0)$, the evaluation results were similar as in the case of range $(5.0, 8.0)$.

The JAS sample is used to estimate the population size of each post-stratum. We consider the sample size ratio, $n_J/n_{\mathcal{L}} = 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0$, where n_J is the total sample size in June and $n_{\mathcal{L}}$ is the total sample size for ALS.

Simulation procedure.

Recall that we have assumed both JAS and ALS have the same sampling frame for the purpose of this simulation study. The simulation procedure and the computations made are as follows:

(1) For a given stratum h, a sample (corresponding to JAS) of size $n_{Jh} = n_{\mathcal{L}h} * r$, $(r = 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0)$ is selected from the population. The sample data are used to compute the estimate of the k-th post-stratum population count, $\hat{M}_k$, for $k = 1, 2, ..., K$.

(2) For a given stratum h, an independent sample (corresponding to ALS) of size $n_{\mathcal{L}h}$, is selected from the ALS. The sample data are used to compute $\hat{\bar{Y}}_k$ or $\bar{y}_k$, the sample average of k-th post-stratum, $k = 1, 2, ..., K$.

(3) For each sample selected according to the parameters chosen, we compute $\hat{Y}_{G\,unwt}$, $\hat{Y}_{G\,unwt(C)}$, $\hat{Y}_{G\,wt}$ and the direct expansion estimator $\hat{Y}_{st}$.

(4) Repeat the above step 2,000 times and compute its mean deviation from the true population value as an estimate for the bias. The percent biases of the four estimators, $\hat{Y}_{G\,unwt}$, $\hat{Y}_{G\,unwt(C)}$, $\hat{Y}_{G\,wt}$ and $\hat{Y}_{st}$ are computed as follows:

$$B_{st} = \left[\left(\frac{1}{2000}\sum_{2000\text{ times}} \hat{Y}_{st}\right)/Y - 1\right] * 100\%,$$

$$B_{u} = \left[\left(\frac{1}{2000}\sum_{2000\text{ times}} \hat{Y}_{G\,unwt}\right)/Y - 1\right] * 100\%,$$

$$B_{u(C)} = \left[\left(\frac{1}{2000}\sum_{2000\text{ times}} \hat{Y}_{G\,unwt(C)}\right)/Y - 1\right] * 100\%,$$

$$B_{w} = \left[\left(\frac{1}{2000}\sum_{2000\text{ times}} \hat{Y}_{G\,wt}\right)/Y - 1\right] * 100\%,$$

where $Y = \sum_{h=1}^{H} N_{Jh}\bar{Y}_h$ is the population total.

(5) Similarly, compute its mean squared deviations from the true population value as an estimate for the MSE:

$$V = \frac{1}{2000}\left(\sum_{2000\text{ times}} (\hat{Y}_{st} - Y)^2\right),$$

$$S_{u} = \frac{1}{2000}\left(\sum_{2000\text{ times}} (\hat{Y}_{G\,unwt} - Y)^2\right),$$

$$S_{u(C)} = \frac{1}{2000}\left(\sum_{2000\text{ times}} (\hat{Y}_{G\,unwt(C)} - Y)^2\right),$$

$$S_{w} = \frac{1}{2000}\left(\sum_{2000\text{ times}} (\hat{Y}_{G\,wt} - Y)^2\right).$$

(6) The sample variance, S_A, computed according to the approximate variance formula given in the previous section (with the necessary simplification) is also obtained for each iteration and then is averaged over all 2000 iterations.

(7) The relative percentage biases, B_u, $B_{u(C)}$, and B_w, and the relative efficiencies, V/S_u, $V/S_{u(C)}$, and V/S_w, of the post-stratified estimators are computed.

In order to distinguish between various parameter inputs and stratifications, we also compute the design efficiency for the original stratification and for the post-stratification as follows:

$$\text{Deff-H} = \frac{\frac{1-f}{n_{\mathcal{L}}} S^2}{\sum_{h=1}^{H} \frac{1-f_h}{n_{\mathcal{L}h}} W_h^2 S_h^2},$$

where $n_{\mathcal{L}}$, the total sample size, is defined in Equation (29),

$$f = \frac{\sum_{h=1}^{H} n_{\mathcal{L}h}}{\sum_{h=1}^{H} N_{Jh}}, \quad f_h = \frac{n_{\mathcal{L}h}}{N_{Jh}}, \quad W_h = \frac{N_{Jh}}{\sum_{h=1}^{H} N_{Jh}}$$

and

$$S_h^2 = \frac{1}{N_{Jh}-1} \sum_{i=1}^{N_{Jh}} (y_{hi} - \bar{Y}_h)^2, \quad S^2 = \frac{1}{\sum_{h=1}^{H} N_{Jh} - 1} \sum_{h=1}^{H} \sum_{i=1}^{N_{Jh}} (y_{hi} - \bar{Y})^2.$$

Similarly, the design efficiency of the post-stratification, Deff-K, is computed. The design efficiency, Deff-H (or Deff-K) is the relative efficiency that can be expected for the stratified (or post-stratified) estimator over the direct expansion estimator when the population counts are known. Of course, the population counts for post-stratification are unknown in our case and these have to be estimated from JAS.

Numerical Results.

The simulation evaluation results are listed in Tables 1–4. In two cases (Tables 1–2), Deff-H and Deff-K are not too much different and hence the post-stratification is not much more efficient than the original stratification. However, Deff-K is substantially higher than Deff-H in the other two cases (Tables 3–4), making the post-stratification much more efficient than the original stratification. Based on the numerical results, the following conclusions are drawn:

(1) The relative efficiencies for the three post-stratified estimators, the weighted ($\hat{Y}_{G\,wt}$), the unweighted ($\hat{Y}_{G\,unwt}$), and the unweighted "combined" ($\hat{Y}_{G\,unwt(C)}$),

as shown in Tables 1–4 are increasing functions of $n_J/n_{\mathcal{L}}$, the ratio of sample sizes between JAS and ALS. This can be attributed to the fact that the larger the sample size is in JAS, more efficiently the post-stratum population counts are estimated. One also finds that when the two sample sizes n_J and $n_{\mathcal{L}}$ are about equal, there is no gain in the post-stratified estimators over the direct expansion estimator.

(2) When the total sample size in ALS is small, say $n_{\mathcal{L}} = 50$, and when the post-stratification is not effective as reflected in the Deff-K value of being close to 1, all three post-stratified estimators have larger variances than the direct expansion estimator $\hat{Y}_{st}$ [Tables 1 and 2 with $n_{\mathcal{L}} = 50$].

(3) For a moderate to large ALS sample size, say $n_{\mathcal{L}} = 100, 200$ or 400, the combined unweighted estimators ($\hat{Y}_{G\,unwt(C)}$) and the weighted estimator ($\hat{Y}_{G\,wt}$) have smaller variances than the direct expansion estimator ($\hat{Y}_{st}$), especially when the post-stratification is more effective. [Tables 1–4]. On the other hand, when the efficiencies of the original and post-stratifications are about the same [Tables 1 and 2], $\hat{Y}_{G\,unwt(C)}$ may not be better than $\hat{Y}_{st}$.

(4) The observed variance approximation S_A is very close to the observed mean square error (MSE) of the post-stratified estimator $\hat{Y}_{G\,wt}$ when $n_{\mathcal{L}} \geq 100$. However, when the ALS sample size is small, say $n_{\mathcal{L}} = 50$, the approximate variance formula underestimates the observed MSE by 30%-50%.

(5) Among three post-stratified estimators considered, the weighted estimator ($\hat{Y}_{G\,wt}$) tends to have the smallest bias, and the unweighted estimator ($\hat{Y}_{G\,unwt}$), the largest bias. For example, when $n_{\mathcal{L}} \geq 100$, the bias of $\hat{Y}_{G\,wt}$ is always less than 0.16% of the true population total; whereas the largest biases of the unweighted estimators ($\hat{Y}_{G\,unwt(C)}$ and $\hat{Y}_{G\,unwt}$) are about 0.6% and 1.0% of the true population total. Of course, the direct expansion estimator $\hat{Y}_{st}$ is an unbiased estimator of the population total.

(6) The unweighted "separate" estimator ($\hat{Y}_{G\,unwt}$) is the worst post-stratified estimator both in terms of the bias and the MSE. The unweighted "combine" estimator ($\hat{Y}_{G\,unwt(C)}$) is slightly better than the weighted estimator ($\hat{Y}_{G\,wt}$) in some cases [Tables 3 and 4], and $\hat{Y}_{G\,wt}$ is slightly better $\hat{Y}_{G\,unwt(C)}$ in one case [Table 1]; while in other case [Table 2], $\hat{Y}_{G\,unwt(C)}$ is better when $n_{\mathcal{L}} \leq 200$ and $\hat{Y}_{G\,wt}$ is better when $n_{\mathcal{L}} = 400$. The main reason is that as the sample size becomes larger, the bias (which is of the constant order) becomes more dominant than the variance (which is of order $1/n_{\mathcal{L}}$) in the MSE and thus the weighted estimator ($\hat{Y}_{G\,wt}$) becomes more efficient because it has smaller bias.

RECOMMENDATIONS

(1) When the sample size $n_{\mathcal{L}}$ is small, the post-stratification estimators are not recommended even when there is a good post-stratification variable.

(2) The general weighted post-stratified estimators ($\hat{Y}_{G\,wt}$) is recommended provided $n_{\mathcal{L}}$ is moderate to large and there is an effective post-stratification variable. The unweighted "combined" estimator $\hat{Y}_{G\,unwt(C)}$ sometimes tends to have a slightly smaller variance, however, $\hat{Y}_{G\,wt}$ tends to have a smaller bias. There is one major advantage of $\hat{Y}_{G\,wt}$ over $\hat{Y}_{G\,unwt(C)}$: the variance formula and its variance estimate are readily available for $\hat{Y}_{G\,wt}$. Therefore, $\hat{Y}_{G\,wt}$ is recommended over $\hat{Y}_{G\,unwt(C)}$.

(3) The performance of the post-stratified estimators is a function of the ratio of two sample sizes, $n_{\mathcal{J}}/n_{\mathcal{L}}$. The relative efficiency of $\hat{Y}_{G\,wt}$ or $\hat{Y}_{G\,unwt(C)}$ over $\hat{Y}_{st}$ increases as $n_{\mathcal{J}}/n_{\mathcal{L}}$ increases. For a post-stratified estimator to be efficient, $n_{\mathcal{J}}/n_{\mathcal{L}} > 1$. Based on the results in Tables 1–4, it is recommended to have the sample size ratio $n_{\mathcal{J}}/n_{\mathcal{L}} \geq 2$.

(4) The Taylor series approximate variance formula is recommended for estimating the design variance of the post-stratified weighted estimator, $\hat{Y}_{G\,wt}$, since it is found to be reasonably accurate for moderate size samples and highly accurate for large size samples.

In the above recommendations, the sample size $n_{\mathcal{L}}$ is termed small, moderate or large on the basis of the following criteria for average number of sample units per stratum, say $\bar{n}_{\mathcal{L}h}$: $n_{\mathcal{L}}$ is small if $\bar{n}_{\mathcal{L}h} < 10$, $n_{\mathcal{L}}$ is moderate if $10 \leq \bar{n}_{\mathcal{L}h} \leq 30$, and $n_{\mathcal{L}}$ is large if $\bar{n}_{\mathcal{L}h} > 30$.

Table 1: Relative Percent Bias, Relative Efficiency and Approximate Variance Ratio.

deff-H= 1.6764, deff-K= 1.8743

$\sigma_3^2 = 1.00$, $\beta \sim U(1.0, 2.0)$

$n_{\mathcal{L}} = 50$

	Rel. Percent Bias			Rel. Efficiency			Ratio
$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.974	0.604	0.153	0.478	0.545	0.503	0.813
1.500	0.846	0.634	-0.001	0.494	0.550	0.511	0.672
2.000	0.963	0.587	0.142	0.549	0.613	0.588	0.607
2.500	0.920	0.637	0.088	0.676	0.738	0.725	0.651
3.000	0.985	0.632	0.147	0.659	0.743	0.727	0.642
4.000	1.090	0.732	0.257	0.750	0.857	0.857	0.635
5.000	0.952	0.582	0.102	0.720	0.799	0.790	0.547

$n_{\mathcal{L}} = 100$

$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.942	0.473	0.080	0.570	0.742	0.711	1.041
1.500	0.978	0.549	0.128	0.662	0.875	0.872	1.010
2.000	0.918	0.458	0.066	0.832	1.140	1.109	1.026
2.500	1.022	0.571	0.157	0.839	1.160	1.213	1.005
3.000	0.989	0.527	0.128	0.863	1.236	1.265	1.019
4.000	1.000	0.538	0.136	0.916	1.315	1.376	0.975
5.000	0.969	0.503	0.103	0.991	1.457	1.504	1.010

$n_{\mathcal{L}} = 200$

$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.909	0.411	0.028	0.507	0.774	0.772	1.013
1.500	0.904	0.401	0.027	0.629	1.020	1.061	1.040
2.000	0.999	0.482	0.119	0.594	1.020	1.133	0.973
2.500	0.946	0.446	0.068	0.674	1.179	1.304	1.009
3.000	0.915	0.405	0.027	0.653	1.167	1.258	0.975
4.000	0.957	0.443	0.070	0.699	1.316	1.500	1.017
5.000	0.946	0.448	0.062	0.719	1.340	1.580	1.001

$n_{\mathcal{L}} = 400$

$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.921	0.417	0.011	0.388	0.729	0.858	0.993
1.500	0.921	0.416	0.018	0.405	0.837	1.045	1.013
2.000	0.931	0.433	0.023	0.421	0.910	1.165	0.974
2.500	0.951	0.451	0.049	0.416	0.928	1.279	0.984
3.000	0.945	0.449	0.034	0.454	1.052	1.508	1.023
4.000	0.955	0.460	0.045	0.451	1.064	1.600	1.003
5.000	0.958	0.460	0.052	0.463	1.115	1.724	0.994

Table 2: Relative Percent Bias, Relative Efficiency and Approximate Variance Ratio.

deff-H= 1.2790, deff-K= 1.4369

$\sigma_3^2 = 2.00$, $\beta \sim U(1.0, 2.0)$

$n_{\mathcal{L}} = 50$

	Rel. Percent Bias			Rel. Efficiency			Ratio
$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.929	0.555	0.166	0.614	0.670	0.610	0.843
1.500	0.777	0.564	-0.033	0.611	0.656	0.597	0.702
2.000	0.915	0.537	0.152	0.663	0.708	0.669	0.672
2.500	0.854	0.570	0.069	0.768	0.807	0.759	0.696
3.000	0.916	0.563	0.123	0.779	0.833	0.787	0.718
4.000	1.078	0.720	0.295	0.814	0.875	0.833	0.689
5.000	0.884	0.518	0.080	0.809	0.850	0.816	0.625

$n_{\mathcal{L}} = 100$

$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.891	0.423	0.076	0.726	0.842	0.777	0.983
1.500	0.932	0.502	0.142	0.783	0.914	0.859	0.968
2.000	0.841	0.380	0.043	0.942	1.099	1.013	0.987
2.500	0.961	0.512	0.135	0.948	1.107	1.062	0.944
3.000	0.938	0.474	0.120	0.933	1.113	1.056	0.999
4.000	0.961	0.500	0.138	0.966	1.141	1.095	0.928
5.000	0.895	0.430	0.072	1.039	1.232	1.161	0.959

$n_{\mathcal{L}} = 200$

$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.849	0.351	0.018	0.701	0.898	0.835	0.981
1.500	0.819	0.316	-0.004	0.819	1.055	0.997	0.992
2.000	0.971	0.454	0.139	0.776	1.050	1.032	0.947
2.500	0.902	0.402	0.074	0.860	1.155	1.109	0.962
3.000	0.866	0.356	0.022	0.833	1.113	1.055	0.928
4.000	0.914	0.401	0.070	0.869	1.197	1.158	0.967
5.000	0.889	0.390	0.052	0.890	1.209	1.189	0.976

$n_{\mathcal{L}} = 400$

$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.851	0.347	-0.012	0.598	0.897	0.911	0.990
1.500	0.866	0.362	0.018	0.615	0.968	1.010	1.005
2.000	0.872	0.374	0.010	0.638	1.017	1.037	0.956
2.500	0.917	0.419	0.069	0.619	1.009	1.089	0.973
3.000	0.898	0.402	0.031	0.672	1.122	1.212	1.030
4.000	0.912	0.418	0.051	0.669	1.115	1.214	0.984
5.000	0.918	0.419	0.061	0.677	1.131	1.256	0.974

Table 3: Relative Percent Bias, Relative Efficiency and Approximate Variance Ratio.

deff-H= 1.2635, deff-K= 4.6269

$\sigma_3^2 = 1.00$, $\beta \sim U(3.0, 4.0)$

$n_{\mathcal{L}} = 50$

	Rel. Percent Bias			Rel. Efficiency			Ratio
$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.474	-0.008	-0.008	0.650	0.730	0.657	0.822
1.500	0.340	0.072	-0.163	0.725	0.809	0.723	0.688
2.000	0.489	-0.004	0.003	0.874	0.957	0.891	0.623
2.500	0.424	0.056	-0.070	1.131	1.225	1.144	0.658
3.000	0.492	0.028	-0.008	1.116	1.234	1.140	0.631
4.000	0.622	0.153	0.130	1.407	1.598	1.491	0.630
5.000	0.502	0.019	-0.008	1.321	1.430	1.361	0.522

$n_{\mathcal{L}} = 100$

$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.524	-0.094	0.014	0.771	0.914	0.828	1.016
1.500	0.584	0.017	0.083	1.018	1.245	1.116	0.982
2.000	0.499	-0.105	-0.003	1.380	1.692	1.504	0.987
2.500	0.627	0.033	0.116	1.533	1.994	1.807	1.019
3.000	0.589	-0.020	0.081	1.684	2.218	1.968	0.974
4.000	0.594	-0.015	0.086	1.954	2.626	2.358	0.971
5.000	0.574	-0.040	0.064	2.236	3.060	2.739	1.005

$n_{\mathcal{L}} = 200$

$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.521	-0.139	0.003	0.761	0.957	0.864	1.009
1.500	0.533	-0.132	0.017	1.097	1.440	1.325	1.049
2.000	0.610	-0.072	0.094	1.130	1.646	1.472	0.944
2.500	0.556	-0.105	0.042	1.392	2.002	1.828	0.994
3.000	0.525	-0.152	0.001	1.402	2.009	1.847	0.974
4.000	0.566	-0.113	0.044	1.677	2.662	2.445	1.006
5.000	0.568	-0.096	0.048	1.821	2.977	2.772	0.989

$n_{\mathcal{L}} = 400$

$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.540	-0.126	0.004	0.709	0.997	0.937	0.985
1.500	0.531	-0.137	-0.001	0.853	1.292	1.232	1.001
2.000	0.552	-0.108	0.017	0.997	1.706	1.567	0.986
2.500	0.562	-0.100	0.032	1.048	1.924	1.782	0.987
3.000	0.566	-0.092	0.029	1.192	2.352	2.171	0.982
4.000	0.570	-0.085	0.032	1.286	2.768	2.568	0.983
5.000	0.574	-0.085	0.040	1.383	3.241	3.030	1.011

Table 4: Relative Percent Bias, Relative Efficiency and Approximate Variance Ratio.

deff-H= 1.1857, deff-K= 2.9776

$\sigma_3^2 = 2.00$, $\beta \sim U(3.0, 4.0)$

$n_{\mathcal{L}} = 50$

	Rel. Percent Bias			Rel. Efficiency			Ratio
$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.447	-0.038	0.000	0.690	0.760	0.682	0.838
1.500	0.298	0.029	-0.182	0.726	0.794	0.709	0.689
2.000	0.459	-0.034	0.009	0.864	0.925	0.864	0.651
2.500	0.384	0.015	-0.082	1.079	1.145	1.059	0.683
3.000	0.450	-0.013	-0.022	1.082	1.163	1.072	0.677
4.000	0.614	0.146	0.153	1.277	1.396	1.292	0.670
5.000	0.461	-0.021	-0.021	1.226	1.285	1.224	0.583

$n_{\mathcal{L}} = 100$

$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.493	-0.125	0.012	0.807	0.922	0.837	1.008
1.500	0.555	-0.012	0.091	0.997	1.161	1.036	0.964
2.000	0.452	-0.152	-0.017	1.317	1.502	1.349	0.977
2.500	0.590	-0.004	0.103	1.428	1.684	1.538	0.982
3.000	0.557	-0.052	0.077	1.492	1.785	1.597	0.976
4.000	0.571	-0.039	0.087	1.638	1.956	1.774	0.944
5.000	0.530	-0.084	0.045	1.845	2.193	1.983	0.973

$n_{\mathcal{L}} = 200$

$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.485	-0.175	-0.004	0.811	0.961	0.868	0.995
1.500	0.481	-0.184	-0.002	1.113	1.321	1.222	1.019
2.000	0.593	-0.089	0.106	1.125	1.468	1.319	0.936
2.500	0.529	-0.132	0.045	1.343	1.707	1.547	0.974
3.000	0.495	-0.181	-0.002	1.307	1.625	1.491	0.943
4.000	0.540	-0.139	0.044	1.513	2.003	1.830	0.981
5.000	0.532	-0.131	0.042	1.621	2.133	1.987	0.974

$n_{\mathcal{L}} = 400$

$n_J/n_{\mathcal{L}}$	B_u	$B_{u(C)}$	B_w	V/S_u	$V/S_{u(C)}$	V/S_w	S_A/S_w
1.000	0.498	-0.169	-0.010	0.782	0.997	0.946	0.988
1.500	0.498	-0.170	-0.001	0.911	1.217	1.168	0.996
2.000	0.516	-0.144	0.009	1.032	1.471	1.357	0.965
2.500	0.541	-0.120	0.044	1.064	1.608	1.490	0.971
3.000	0.537	-0.121	0.027	1.221	1.918	1.785	1.009
4.000	0.543	-0.111	0.036	1.278	2.073	1.923	0.977
5.000	0.550	-0.109	0.046	1.349	2.248	2.114	0.979

APPENDIX A:

ESTIMATING THE DESIGN-BIAS AND ITS VARIANCE

Since any direct expansion estimator is design unbiased, the design-bias of the generalized post-stratified estimator,$\hat{G}$, can be estimated by:

$$\hat{\beta} = \hat{G} - \hat{Y}^d = \sum_{k=1}^{K} \left(\hat{G}_k - \hat{Y}_k^d \right) = \sum_{k=1}^{K} \left[\left(\frac{\hat{Z}_k \hat{Y}_k}{\hat{X}_k} \right) - \hat{Y}_k^d \right] \tag{A1}$$

where $\hat{Y}_k^d$ represents the direct expansion estimator of the total,Y_k, for the item of interest for post-stratum k derived from the Agricultural Labor Survey, and where $\hat{Z}_k$, $\hat{X}_k$ and $\hat{Y}_k$ are defined as in Equation (1).

In terms of the above notation, the first-order approximation of the design-variance of the estimated bias, $\hat{\beta}$, of the generalized post-stratified estimator estimator, $\hat{G}$, can be written as:

$$V(\hat{\beta}) = V \left(\sum_{k=1}^{K} \partial G_{Z_k} \hat{Z}_k + \partial G_{Y_k} \hat{Y}_k + \partial G_{X_k} \hat{X}_k - \hat{Y}_k^d \right) \tag{A2}$$

where the partial derivatives and other notation are interpreted as in Equation (2). Equation (A2) is easily justified by considering separately two cases, one where $\hat{Y}_k = \hat{Y}_k^d$, and the other where $\hat{Y}_k \neq \hat{Y}_k^d$, and then observing that $\partial Y_{Y_k^d}^d = 1$, for $k = 1, \cdots, K$. (Actually for completeness, the proof requires one to consider, as a special case, the special situation where $\hat{Y}_k \neq \hat{Y}_k^d$ and yet $\hat{Y}_{k \cap D} = \hat{Y}_{k \cap D}^d$ for some subdomain D. An example of this situation occurs when $\hat{Y}_k^d$ is the direct expansion estimate over both the list and NOL, and $\hat{Y}_k$ is the direct expansion estimate over the list only.)

Since Equation (A2) can be derived from Equation (2) by replacing $\partial G_{Y_k} \hat{Y}_k$ with $\partial G_{Y_k} \hat{Y}_k - \hat{Y}_k^d$, it follows that the computations necessary to produce the first-order Taylor series variance estimate of design-bias of the generalized post-stratified estimator estimator, $\hat{G}$, are given by the equation that result from making similar changes to Equation (16):

$$V(\hat{\beta}) = \sum_{h \in H_{L_J}} V_{L_J} \left(\sum_{i \in n_{L_J h}} t_{L_J hi} \right) + \sum_{h \in H_{L_{\mathcal{L}}}} V_{L_{\mathcal{L}}} \left(\sum_{i \in n_{L_{\mathcal{L}} h}} t_{L_{\mathcal{L}} hi} \right) +$$

$$\sum_{h \in H_{A_{J \setminus \mathcal{L}}}} V_{A_{J \setminus \mathcal{L}}} \left(\sum_{i \in n_{A_{J \setminus \mathcal{L}} h}} t_{A_{J \setminus \mathcal{L}} hi} \right) + \sum_{h \in H_{A_{\mathcal{L}}}} V_{A_{\mathcal{L}}} \left(\sum_{i \in n_{A_{\mathcal{L}} h}} (t_{A_{J \cap \mathcal{L}} hi} + \hat{t}_{A_{\mathcal{L}} hi}) \right), \quad \text{(A3)}$$

where

$$t_{L_J hi} = W_{L_J h} \sum_{k=1}^{K} \partial G_{Z_k} z_{L_J hi} \, \delta(k, L_J hi),$$

$$t_{L_{\mathcal{L}} hi} = W_{L_{\mathcal{L}} h} \sum_{k=1}^{K} [\partial G_{Y_k} y_{L_{\mathcal{L}} hi} + \partial G_{X_k} x_{L_{\mathcal{L}} hi}] \, \delta(k, L_{\mathcal{L}} hi),$$

$$t_{A_{J \setminus \mathcal{L}} hi} = W_{A_J h} \sum_{j \in M_{A_{J \setminus \mathcal{L}} hi}} \sum_{k=1}^{K} \partial G_{Z_k} z_{A_{J \setminus \mathcal{L}} hij} \, \delta(k, A_{J \setminus \mathcal{L}} hij),$$

$$t_{A_{J \cap \mathcal{L}} hi} = W_{A_J h} \sum_{j \in M_{A_{J \cap \mathcal{L}} hi}} \sum_{k=1}^{K} \partial G_{Z_k} z_{A_{J \cap \mathcal{L}} hij} \, \delta(k, A_{J \cap \mathcal{L}} hij),$$

$$\hat{t}_{A_{\mathcal{L}} hi} = W_{A_{\mathcal{L}} h} \sum_{j \in m_{A_{\mathcal{L}} hi}} W_{sA_{\mathcal{L}} hij} \sum_{k=1}^{K} \left[\partial G_{Y_k} y_{A_{\mathcal{L}} hij} - y^{d}_{A_{\mathcal{L}} hij} + \partial G_{X_k} x_{A_{\mathcal{L}} hij} \right] \delta(k, A_{\mathcal{L}} hij).$$

The stratum level variance estimates in Equation (A3) are computed with the formula given in Equation (17).

APPENDIX B:
VARIANCE OF THE RATIO OF POST-STRATIFIED ESTIMATORS FROM TWO OCCASIONS

Suppose a population has been sampled and the characteristics of interest estimated with the same generalized post-stratified estimators on two occasions. (For example, a characteristic on the Agricultural Labor Survey has been estimated from the common

replicates on two monthly surveys.) Let

$$\hat{R} = \frac{\hat{G}^{t_1}}{\hat{G}^{t_2}} \tag{B1}$$

where

$$\hat{G}^{t_1} = \sum_{k=1}^{K} \hat{G}_k^{t_1} = \sum_{k=1}^{K} \frac{\hat{Z}_k \hat{Y}_k^{t_1}}{\hat{X}_k}$$

and

$$\hat{G}^{t_2} = \sum_{k=1}^{K} \hat{G}_k^{t_2} = \sum_{k=1}^{K} \frac{\hat{Z}_k \hat{Y}_k^{t_2}}{\hat{X}_k}.$$

The superscripts t_1 and t_2 represent respectively estimates from the first and second occasion. The Taylor series estimator of the variance of the estimated ratio is:

$$V(\hat{R}) = \hat{R}^2 \left(\frac{V(\hat{G}^{t_1})}{(\hat{G}^{t_1})^2} + \frac{V(\hat{G}^{t_2})}{(\hat{G}^{t_2})^2} - 2\frac{COV(\hat{G}^{t_1}, \hat{G}^{t_2})}{\hat{G}^{t_1}\hat{G}^{t_2}} \right) \tag{B2}$$

Since

$$\hat{G}^{t_1} + \hat{G}^{t_2} = \sum_{k=1}^{K} \left(\frac{\hat{Z}_k \hat{Y}_k^{t_1}}{\hat{X}_k} + \frac{\hat{Z}_k \hat{Y}_k^{t_2}}{\hat{X}_k} \right) = \sum_{k=1}^{K} \frac{\hat{Z}_k \left(\hat{Y}_k^{t_1} + \hat{Y}_k^{t_2} \right)}{\hat{X}_k}$$

and

$$2COV(\hat{G}^{t_1}, \hat{G}^{t_2}) = V\left(\hat{G}^{t_1} + \hat{G}^{t_2}\right) - V\left(\hat{G}^{t_1}\right) - V\left(\hat{G}^{t_2}\right),$$

the variance of the ratio is easily evaluated by applying the computational variance formula given in Equation (16) to each of the components $\hat{G}^{t_1}, \hat{G}^{t_2}$ and $\hat{G}^{t_1} + \hat{G}^{t_2}$, and then substituting the results in Equation (B2).

The derivation given above for the variance of the ratio of two generalized post-stratified estimators simply combines two Taylor series variance formulas. Thus, the derived formula is not the same as one would have obtained by applying the Taylor series methodology directly to the ratio of the two generalized post-stratified estimators, which can be easily verified. However, the two variance formulas are asymptotically equivalent.

REFERENCES

Cochran, W. G. (1977), *Sampling Techniques*, 3rd edition, John Wiley & Sons, New York.

Holt, D. and Smith, T.M.F. (1979), "Post-Stratification," *Journal of Royal Statistical Society* Series A, **142**, p. 33-46.

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

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

Rumburg, S., Perry, C. R., Chhikara, R. S., and Iwig, W. C. (1993), "Analysis of Generalized Post-stratification Approach for the Agricultural Labor Survey," SRB Research Report No. SRB-93-05, Washington, D. C., July 1993.

Särndal, C.E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

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

Wolter, K. M. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.