BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-84/07

# METHODOLOGY FOR OPTIMAL DUAL FRAME SAMPLE DESIGN

by

Paul P. Biemer

U.S. Bureau of the Census
Washington, D.C. 20233

This series contains research reports, written by or in cooperation with staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233.

Recommended by: Kirk M. Wolter

Report completed: August 15, 1983

Report issued: January 24, 1984

# METHODOLOGY FOR OPTIMAL DUAL FRAME SAMPLE DESIGN

by

Paul P. Biemer

Paul P. Biemer is a Principal Statistical Researcher in the Center for Survey Methods Research at the Bureau of the Census, Washington, D.C.

# METHODOLOGY FOR OPTIMAL DUAL FRAME SAMPLE DESIGN

by

Paul Biemer

## 1. INTRODUCTION

Hartley (1974) defines a multiple frame survey as "a set of several (single frame) surveys whose samples are combined to provide parameter estimates for the union of frames." This general methodology finds great application in the case where there are two frames involved--one being a telephone frame and the other an areal frame. The "dual frame" survey has an advantage over the single frame telephone sample survey in that it offers complete coverage of the population and, therefore, can provide unbiased estimators of population parameters. This, of course, comes with a higner cost for data collection. In most cases, this cost is less than the cost of a single frame area sample survey which also offers full coverage. A number of authors have demonstrated this through simulation (see, for example, Hartley 1962; Lund 1968; and Casady and Sirken 1980).

The potential advantages of low costs without loss of population coverage are especially attractive to government data collection agencies who report characteristics, such as crime, unemployment, and health since these statistics can be substantially affected by the omission of non-telephone domains (cf., Thornberry and Massey 1978 and McGowan 1981). Recently, the Bureau of the Census has initiated a considerable program of research and development to investigate the many issues surrounding this new methodology. Among the major sampling topics under investigation are:

1. sample design strategies for the allocation of resources to the two frames in order to minimize cost and error,

ABSTRACT

This paper provides a comprehensive and systematic framework for evaluating a wide range of statistical information in order to optimize the design of dual frame surveys. Formulae for the total mean square error of dual frame survey estimators are derived for general stratified multistage sample designs under a model which incorporates nonsampling bias and variance terms. A general procedure for selecting resource allocations which minimize the total survey error for a fixed budget is developed. The methodology is applied to the Current Population Survey for which a number of uses of the procedure are illustrated.

KEYWORDS: Multiple frame surveys; nonsampling or response error; complex surveys.

2. the estimation of population target parameters from dual frame samples,
3. procedures for minimizing the impact on estimates and costs of the conversion from an areal frame to a dual frame survey, and
4. the estimation and evaluation of nonsampling errors as they affect the accuracy of the estimators and the allocation of resources.

Hartley (1962) addressed items (1) and (2) for simple random sampling in each frame and, subsequently, a number of authors have offered improved estimators. An important paper by Casady and Sirken suggested applying the multiple frame metholodgy to telephone surveys; Casady, Snowden, and Sirken (1981) consider a dual frame telephone survey for the National Health Interview Survey. However, it is clear that a more comprehensive development of the methodology for complex surveys is needed in order to handle most of the sampling problems encountered in practical dual frame survey design.

This paper provides the methodology for addressing items (1), (2) and (3) above for general stratified multistage survey designs. Since data quality is a key issue in the decision to convert to a telephone/areal dual frame survey, a simple model for studying nonsampling error (item (4)) is proposed. Finally, a general method for dual frame survey optimization is developed and applied to a current survey of the Bureau of the Census.

The types of surveys covered here are essentially general stratified multistage surveys where the last stage units are selected with equal probability within the next-to-last stage units. To simplify the exposition of the results, the estimation and optimization formulae will first be described in terms of two-stage sampling in each frame and will then be outlined for the general multistage situation.

## 2. THE GENERAL SURVEY SPECIFICATIONS

Consider a pair of surveys, referred to as Survey A and Survey B, for estimating the total Y for some characteristic y of a population of M elements. The population elements may be any units that can be uniquely defined within the ultimate sampling units--for example, persons, families, households, etc. within dwelling units or groups of dwelling units. In our discussion, Survey A is an area sample survey and Survey B is a telephone sample survey; however, the methodology can be easily extended to handle any dual frame survey for which the Survey B frame is contained in the Survey A frame.

### 2.1 Description of Survey A

The sampling frame is an areal frame (denoted by Frame A) where the listing units are dwelling units. The sample design is a stratified two-stage design (a condition to be relaxed later) where the secondary sampling units are area segments of dwellings. The segments are selected by an equal probability without replacement selection method (EPSEM) while any equal or unequal probability selection method is possible for the primary units. Interviewer assignments are composed of segments which are randomly selected within a primary. Each interviewer is assigned the same number of segments.

### 2.2 Description of Survey B

The frame contains a list (which may be implicit) of telephone numbers (denoted by Frame B). For simplicity, it is assumed that each population element may be linked to, at most, one telephone number. The sample is selected completely independently of Survey A using a stratified two-stage design (general multistage telephone samples are treated later). The secondary units are telephone numbers sampled with EPSEM without replacement within each primary unit. Interviewer assignments are made up of telephone numbers which are randomly assigned without regard to primary or stratum boundaries.

## 3. MODEL FOR NONSAMPLING ERRORS

To view the impact of nonsampling error on the accuracy of dual frame estimators, an additive error model (also used in more recent literature) is adopted in which the errors made by a particular interviewer are correlated through an additive error term. The study is confined to one particular content item. Further, to justify the subsequent model assumptions, the data are supposed to be quantitative.

Denote the true content item of the tth elementary unit of the sth secondary of the pth primary in stratum h for Survey A by $y_{Ahpst}$ and for Survey B by $y_{Bhpst}$. Denote by $x_{Ahpst}$ and $x_{Bhpst}$ the corresponding recorded content items. For Survey A, let the subscript (h,i) denote the ith interviewer in stratum h. Then for elementary units assigned to interviewer (h,i), it is assumed that

$$x_{Ahpst} = y_{Ahpst} + \alpha_{hi} + \varepsilon_{Ahpst}$$

where $\alpha_{hi}$ is the systematic error contributed by the interviewer and $\varepsilon_{Ahpst}$ is the elementary error associated with the unit. Likewise, it is assumed that

$$x_{Bhpst} = y_{Bhpst} + \beta_i + \varepsilon_{Bhpst}$$

where $\beta_i$ is the systematic error associated with the ith interviewer for Survey B and $\varepsilon_{Bhpst}$ the elementary error associated with the unit.

Now, let $D_1$ refer to the elements in the population which belong only to Frame A and let $D_2$ denote the elements belonging to both frames. It is assumed that $\alpha_{hi}$, $\beta_i$, $\varepsilon_{Ahpst}$ and $\varepsilon_{Bhpst}$ are random samples from infinite populations with $E(\alpha_{hi}) = E(\beta_i) = 0$, $V(\alpha_{hi}) = \sigma^2_{\alpha h}$, $V(\beta_i) = \sigma^2_{\beta}$, $E(\varepsilon_{Ahpst}|D_1) = B_{1h}$, $E(\varepsilon_{Ahpst}|D_2) = B_{2h}$ and $E(\varepsilon_{Bhpst}) = B_{Bh}$ where $E(\cdot|D_1)$ and $E(\cdot|D_2)$ denotes restriction to $D_1$ or $D_2$. Furthermore, assume that for a given primary (h, p) in each survey,

$$V'(y_{Ahpst} + \varepsilon_{Ahpst}|D_1) = \sigma^2_{1hp},$$

$$V'(y_{Ahpst} + \varepsilon_{Ahpst}|D_2) = \sigma^2_{2hp},$$

and

$$V'(y_{Bhpst} + \varepsilon_{Bhpst}) = \sigma^2_{Bhp}$$

where $V'$ denotes variance over simple random samples of one element drawn from the populations implied by the contents of the parentheses.

## 4. ESTIMATORS OF THE POPULATION TOTAL

Two classes of general estimators are considered. The first is the estimator proposed by Hartley (1974) which combines domain estimators over all strata before weighting together to form the population estimator. This estimator is applicable for any dual frame design and will be referred to as the "combined" estimator. The second estimator is only appropriate when the strata for Survey A and Survey B are the same or at least coincide so that the strata for one are nested within the strata for the other. This estimator, suggested by Bosecker and Ford (1976), will be referred to as the "separate" estimator. First, a number of symbols must be defined.

4.1 <u>Notation</u>

Let

$m_{\ell hp}$ = number of sample elements belonging to $D_\ell$ ($\ell$ = 1,2) in primary (h,p) for Survey A,

$m_{Ahp} = m_{1hp} + m_{2hp}$,

$m_{Bhp}$ = number of sample elements in primary (h,p) for Survey B,

$\bar{x}_{\ell hp} = \sum\sum\sum_{(p,s,t)\,\varepsilon\, D_\ell} x_{Ahpst}/m_{\ell hp}$, $\ell = 1,2$,

$\bar{x}_{Ahp} = \sum_p \sum_s \sum_t x_{Ahpst}/m_{Ahp}$,

$\bar{x}_{Bhp} = \sum_p \sum_s \sum_t x_{Bhpst}/m_{Bhp}$,

$M_{\ell hp}$ = total number of elements in $D_\ell$ ($\ell$ = 1,2) in primary (h,p) for Survey A,

$M_{Ahp} = M_{1hp} + M_{2hp}$,

$M_{Bhp}$ = total number of elements in primary (h,p) for Survey B,

$q_{Ah}$ = expected number of elements for a Survey A interviewer assignment in stratum h,

$q_B$ = expected number of elements in a Survey B interviewer assignment,

$k_{hp}$ = number of interviewers working in primary (h,p) for Survey A,

$k_B$ = number of interviewers available for Survey B,

$\Pi_A(\Pi_B)$ = set of sampled primary units for Survey A (Survey B),

$\overline{m}_{Ahp} = E(m_{Ahp} | \Pi_A)$,

$\overline{m}_{\ell hp} = M_{\ell hp} \, \overline{m}_{Ahp} / M_{Ahp} = E(m_{\ell hp} | \Pi_A)$, $\ell = 1,2$,

$L_A(L_B)$ = number of strata defined for Survey A (Survey B),

$M_{\ell h}$ = number of elements in $D_\ell$ for the stratum h for Survey A, $\ell = 1,2$,

$M_{Bh}$ = number of elements in stratum h for Survey B,

$\overline{m}_{Ah} = \sum_{p}^{n_{Ah}} \overline{m}_{Ahp}$,

$\overline{m}_{Bh} = \sum_{p}^{n_{Bh}} \overline{m}_{Bhp}$,

$M_{Ah} = M_{1h} + M_{2h}$,

$n_{Ah}(n_{Bh})$ = number of sample primaries in stratum h for Survey A (Survey B),

$N_{Ah}(N_{Bh})$ = number of primaries in stratum h for Survey A (Survey B),

$\theta_h$ or $\theta$ = dual frame survey weight, a constant between 0 and 1 to be optimized,

$\overline{Y}_{\ell hp}$ = true population mean per element for $D_\ell$ ($\ell = 1,2$) in primary (h,p),

$\overline{Y}_{Bhp}$ = true population mean per element for primary (h,p) Survey B,

$B_{\ell h} = \overline{X}_{\ell hp} - \overline{Y}_{\ell hp}$, $\ell = 1,2$,

$B_{Bh} = \overline{X}_{Bhp} - \overline{Y}_{Bhp}$,

## 4.2 The Combined Estimator

Define a new variable $u_{hpst}(\theta)$ for Survey A such that

$$u_{hpst}(\theta) = x_{Ahpst} \text{ if } (h,p,s,t) \in D_1$$
$$= \theta \, x_{Ahpst} \text{ if } (h,p,s,t) \in D_2.$$

Since sampling within primaries is with equal probability, we consider estimators of the form

$$\hat{X}_C = \sum_h^{L_A} \sum_p^{n_{Ah}} w_{Ahp} \, \overline{u}_{hp}(\theta) + (1-\theta) \sum_h^{L_B} \sum_p^{n_{Bh}} w_{Bhp} \, \overline{x}_{Bhp} \qquad (4.2.1)$$

where $\overline{u}_{hp}(\theta) = \sum_s \sum_t u_{hpst}(\theta)/m_{Ahp}$ and $w_{Ahp}$, $w_{Bhp}$ are the usual single frame sample weights which may depend upon the sets of sampled primary units, $\Pi_A$ and $\Pi_B$. This estimator will be referred to as the combined estimator since strata estimators in each frame are combined before weighting by the parameter $\theta$.

Let $E_2$ denote conditional expectation with respect to the nonsampling error distributions as well as within primary sampling given the primary samples $\Pi_A$ and $\Pi_B$ and let $E_1$ denote the expectation over all possible $\Pi_A$ $\Pi_B$. Define the variance operators $V_1$ and $V_2$ analogously. Ignoring the technical bias in $\hat{X}_C$, then $E_2 \, \overline{u}_{hp}(\theta)$ is given by

$$\overline{U}_{hp}(\theta) = \tau_{1hp} \overline{X}_{1hp} + \theta \, \tau_{2hp} \overline{X}_{2hp} \qquad (4.2.2)$$

where $\tau_{1hp} = M_{1hp}/M_{Ahp}$ and $\tau_{2hp} = 1 - \tau_{1hp} = M_{2hp}/M_{Ahp}$.

The total nonsampling bias in $\hat{X}_C$, $B(\hat{X}_C) = E(\hat{X}_C) - Y$, is, therefore,

$$B(\hat{X}_C) = \sum_h^{L_A} M_{Ah} \, b_{Ah}(\theta) + (1-\theta) \sum_h^{L_B} M_{Bh} \, B_{Bh} \qquad (4.2.3)$$

where $b_{Ah}(\theta) = \tau_{1n} B_{1h} + \theta \, \tau_{2h} B_{2h}$.

The variance of $\hat{X}_C$ can be decomposed into between and within primary components using the identity

$$V(\hat{X}_C) = V_1 E_2 \, \hat{X}_C + E_1 V_2 \, \hat{X}_C.$$

Considering the between primary components, we have

$$V_1 E_2(\hat{X}_C) = \sum_h^{L_A} V_1 \sum_p^{n_{Ah}} w_{Ahp}\, \overline{U}_{hp}(\theta) + (1-\theta)^2 \sum_h^{L_B} V_1 \sum_p^{n_{Bh}} w_{Bhp}\, \overline{X}_{Bhp}. \qquad (4.2.4)$$

The analytic form of these components depends upon the primary stage sample selection scheme in each survey. Formulas are provided by classical sampling theory (without nonsampling error), treating the primary stage as the last stage with observations $\overline{U}_{hp}$ and $\overline{X}_{Bhp}$. Special cases of these formulas are considered in the next section.

The within primary variance component of $\hat{X}_C$ depends on the terms

$$\sum_p w_{Ahp}^2\, V_2\, \overline{u}_{hp}(\theta)$$

in Survey A and

$$\sum_p w_{Bhp}^2\, V_2\, \overline{x}_{Bhp} + \sum_{(h,p)} \sum_{(h',p')} w_{Bhp}\, w_{Bh'p'}\, Cov_2(\overline{x}_{Bhp}, \overline{x}_{Bh'p'})$$

in Survey B where $Cov_2$ is the conditional covariance operator analagous to $E_2$ and $V_2$.

Writing $\overline{u}_{hp}(\theta)$ as $t_{1hp}\, \overline{x}_{1hp} + \theta\, t_{2hp}\, \overline{x}_{2hp}$ where $t_{1hp} = m_{1hp}/m_{Ahp}$ and $t_{2hp} = 1 - t_{1hp}$, it follows that, to terms of order $1/\overline{m}_{Ahp}$,

$$V_2\, \overline{u}_{hp}(\theta) \doteq \tau_{1hp}^2\, V_2\, \overline{x}_{1hp} + \theta^2\, \tau_{2hp}^2\, V_2\, \overline{x}_{2hp}$$

$$- 2\, \tau_{1hp}\, \tau_{2hp}\, Cov_2(\overline{x}_{1hp}, \overline{x}_{2hp}) + (\overline{X}_{1hp} - \theta\, \overline{X}_{2hp})^2\, V_2\, (t_{2hp}). \qquad (4.2.5)$$

From Section 3, $\overline{x}_{1hp}$ can be written as

$$\overline{x}_{1hp} = (\overline{y}_{1hp} + \overline{\epsilon}_{1hp}) + \sum_i \nu_1(i;h,p)\, \alpha_{hi}$$

where $\overline{y}_{1hp}$ and $\overline{\epsilon}_{1hp}$ are defined in analogy to $\overline{x}_{1hp}$ and $\nu_1(i;h,p)$, a random variable, is the fraction of $D_1$ sample elements assigned to the ith interviewer in primary (h,p). Approximating $E_2\, \nu_1^2(i;h,p)$ by $1/k_{hp}^2$, we have

$$V_2\,\bar{x}_{1hp} \doteq \delta_{1hp}\,\sigma^2_{1hp}/\bar{m}_{1hp} + \sigma^2_{\alpha h}/k_{hp} \qquad (4.2.6)$$

where $\delta_{1hp} = V_2(\bar{y}_{1hp} + \bar{\varepsilon}_{1hp})\bar{m}_{1hp}/\sigma^2_{1hp}$ is the within primary design effect for $D_1$. (The analytic form of $\delta_{1hp}$ is considered in Appendix A.)

Similarly, we have

$$V_2\,\bar{x}_{2hp} \doteq \delta_{2hp}\,\sigma^2_{2hp}/\bar{m}_{2hp} + \sigma^2_{\alpha h}/k_{hp}$$

and

$$\mathrm{Cov}_2(\bar{x}_{1hp}, \bar{x}_{2hp}) = \rho_{12hp}\left(\frac{\delta_{1hp}\,\delta_{2hp}}{\bar{m}_{1hp}\,\bar{m}_{2hp}}\right)^{\frac{1}{2}} \sigma_{1hp}\,\sigma_{2hp} + \sigma^2_{\alpha h}/k_{hp}$$

where $\rho_{12hp} = \mathrm{Cov}_2(\bar{y}_{1hp}, \bar{y}_{2hp})\,(V_2(\bar{y}_{1hp})\,V_2(\bar{y}_{2hp}))^{-\frac{1}{2}}$ and $\delta_{2hp}$, $\sigma_{2hp}$ are defined similarly as for $\delta_{1hp}$ and $\sigma^2_{1hp}$ in (4.2.6). Finally,

$$V_2\,t_{2hp} \doteq \phi_{Ahp}\,\tau_{1hp}\,\tau_{2hp}/\bar{m}_{Ahp}$$

where $\phi_{Ahp} = \bar{m}_{Ahp}\,V_2\,(t_{2hp})/\tau_{1hp}\,\tau_{2hp}$ is the within primary design effect associated with $t_{2hp}$.

Now, considering the within primary variance component for Survey B, we follow a similar approach to the above; however, now we must consider correlations introduced <u>between</u> primaries and strata as a result of the systematic interviewer errors, $\beta_i$.

Let $\delta_{Bhp}$ denote the within primary design effect for primary (h,p) in Survey B and define

$$\gamma_B = E\,[\sum_h^{L_B} \sum_p^{n_{Bh}} w_{Bhp}\,\nu_B(i;h,p)]^2 \qquad (4.2.7)$$

where, $\nu_B(i;h,p)$ is the fraction of primary (h,p) elements interviewed by the ith interviewer for Survey B.

It can be shown from (4.2.3), (4.2.4) and the above discussion that a general formula for the total mean square error of $\hat{X}_C$ is,

$$MSE(\hat{X}_C) \doteq B^2(\hat{X}_C) + \sum_h^{L_A} \{V_1 \sum_p w_{Ahp} \overline{U}_{hp}(\theta) + E_1 [V_{Awh}(\theta;\Pi_A)]\}$$

$$+ (1-\theta)^2 \sum_h^{L_B} \{V_1 \sum_p w_{Bhp} \overline{X}_{Bhp} + E_1 [V_{Bwh}(\Pi_B)]\} \tag{4.2.8}$$

where

$$V_{Awh}(\theta;\Pi_A) = \sum_p w_{Ahp}^2[\tau_{1hp} \delta_{1hp} \sigma_{1hp}^2 + \theta^2 \tau_{2hp} \delta_{2hp} \sigma_{2hp}^2$$

$$+ (2\theta \tau_{1hp} \tau_{2hp} \delta_{1hp} \delta_{2hp})^{1/2} \sigma_{1hp} \sigma_{2hp} \rho_{12hp}$$

$$+ \phi_{Ahp} \tau_{1hp} \tau_{2hp}(\overline{X}_{1hp} - \theta \overline{X}_{2hp})^2$$

$$+ (\tau_{1hp} + \theta \tau_{2hp})^2 \sigma_{\alpha h}^2 q_{Ah}]/\overline{m}_{Ahp}$$

and

$$V_{Bwh}(\Pi_B) = \sum_p w_{Bhp}^2 \delta_{Bhp} \sigma_{Bhp}^2/\overline{m}_{Bhp} + k_B \gamma_B \sigma_\beta^2/L_B.$$

This result may be compared with Des Raj's (1966) general formulae for determining the variance of an estimator from a multistage survey. Here we have essentially extended his general formulae to cover estimators from complex dual frame surveys and have added components for nonsampling variance and bias.

## 4.3 The Separate Estimator

When strata for one survey are subsets of the strata for the other survey, we call the strata "nested." A stratum which is made up of a number of strata from the companion survey is called a "superstratum." We consider the case where there are L superstrata in the dual frame survey and consider an estimator of the form

$$\hat{X}_S = \sum_{h=1}^{L} [\sum_{p}^{n_{Ah}} w_{Ahp}\, \overline{u}_{hp}(\theta_h) + (1 - \theta_h) \sum_{p}^{n_{Bh}} w_{Bhp}\, \overline{x}_{Bhp}] \tag{4.3.1}$$

called the "separate" estimator. Note here the dual frame weights $\theta_h$ are allowed to vary across superstrata indicating that each superstratum is to be optimized separately with regard to the mix of telephone and area frame sampling. In many cases, this allows the separate estimator to achieve smaller variance then the combined estimator. In (4.3.1), $w_{1hp}$, $w_{2hp}$ and $w_{Bhp}$ are the weights as defined in (4.2.1) where now the summation extends over all p within superstratum h.

Using the results of Section 4.2, it can be shown that $B(\hat{X}_S) = E(\hat{X}_S - Y)$ is given by

$$B(\hat{X}_S) = \sum_{h}^{L} M_{Ah}[b_{Ah}(\theta_h) + (1 - \theta_h)\, \tau_{2h}\, B_{Bh}] \tag{4.3.2}$$

where $b_{Ah}(\theta_h) = b_{Ah}(\theta)$ in (4.2.3) with $\theta$ replaced by $\theta_h$ and $\tau_{2h} = M_{2h}/M_{Ah}$. Now $B_{1h}$, $B_{2h}$ and $B_{Bh}$ are biases which are weighted averages of stratum biases within superstratum h. For example, if Survey A strata are nested within a Survey B stratum, $B_{1h}$ is a combination of Survey A stratum biases weighted by the nested strata weights.

The variance of $\hat{X}_S$ can be obtained directly from previous results. It is

$$\begin{aligned} V(\hat{X}_S) \doteq \sum_{h}^{L} V_1 \sum_{p} w_{Ahp}\, \overline{U}_{hp}(\theta_h) + E_1\, [V_{Awh}(\theta_h; \Pi_A)] \\ + (1 - \theta_h)^2\, \{V_1 \sum_{p} w_{Bhp}\, \overline{X}_{Bhp} + E_1\, [V_{Bwh}(\Pi_B)]\} \end{aligned} \tag{4.3.3}$$

where the variance terms have the same form as in (4.2.7), replacing $\theta$ by $\theta_h$ and extending the summations over all primaries in superstratum h.

### 4.4 Generalization to Multistage Sampling

Certain generalizations are feasible. The role of primary sampling stage can be taken over by any lower stage which we will term the "intermediate stage." The above formulae are restated by calling "primary units," "intermediate units" and "secondary units," "ultimate units." No restrictions must be made on the survey design for stages above the intermediate stage so that unequal or equal probability sampling is permitted for all stages from the primary down to and including the intermediate stage. However, we must still assume that the ultimate units are chosen with EPSEM within the intermediate stage.

## 5. THE SPECIAL CASE OF EPSEM SAMPLING WITHIN EACH STRATUM

Equal probability sampling within each stratum occurs quite frequently in practice. This usually results in considerable simplification in the form of the estimator and its variance. For EPSEM designs, we consider self-weighting estimators satisfying

$$\frac{w_{Ahp}}{m_{Ahp}} = \frac{M_{Ah}}{m_{Ah}}$$

and

$$\frac{w_{Bhp}}{m_{Bhp}} = \frac{M_{Bh}}{m_{Bh}}$$

In this case, the estimators $\hat{X}_C$ and $\hat{X}_S$ simplify to

$$\hat{X}_C = \sum_h^{L_A} M_{Ah}(t_{1h}\,\bar{x}_{1h} + \theta\, t_{2h}\,\bar{x}_{2h}) + (1-\theta) \sum_h^{L_B} M_{Bh}\,\bar{x}_{Bh} \tag{5.1}$$

and

$$\hat{X}_S = \sum_h^{L} M_{Ah}(t_{1h}\,\bar{x}_{1h} + \theta\, t_{2h}\,\bar{x}_{2h} + (1 - \theta_h)\,\tau_{2h}\,\bar{x}_{Bh}) \tag{5.2}$$

where $t_{2h} = m_{2h}/m_{Ah}$ and $t_{1h} = 1-t_{2h}$. In this section, the mean square errors of four estimators which are special cases of (5.1) and (5.2) are examined. These are now described.

For some population characteristics, the telephone domain sizes $\tau_{2h}$ and $M_{Bh}$ are known, for example, from previous Census data. Incorporating this information in the estimators, we have, for the combined form,

$$\hat{X}_{KC} = \sum_{h}^{L_A} M_{Ah}(\tau_{1h}\,\overline{x}_{1h} + \theta_h\,\tau_{2h}\,\overline{x}_{2h}) \tag{5.3}$$

$$+ (1-\theta) \sum_{h}^{L_B} M_{Bh}\,\overline{x}_{Bh},$$

and, for the separate form,

$$\hat{X}_{KS} = \sum_{h}^{L} M_{Ah}(\tau_{1h}\,\overline{x}_{1h} + \theta_h\,\tau_{2h}\,\overline{x}_{2h} + (1-\theta_h)\,\tau_{2h}\,\overline{x}_{Bh}). \tag{5.4}$$

Alternatively, for some populations, the telephone domain sizes may not be known exactly and other estimators of Y may be preferred. It may be possible to use information from "data banks" on the telephone population or combine information from a number of surveys to estimate $\tau_{2h}$ or the $M_{Bh}$. For situations in which no information outside the survey is available, the combined estimator considered is

$$\hat{X}_{UC} = \sum_{h}^{L_A} M_{Ah}(t_{1h}\,\overline{x}_{1h} + \theta\,t_{2h}\,\overline{x}_{2h}) + (1-\theta)\,\hat{M}_B\,\overline{x}_B \tag{5.5}$$

where $\hat{M}_B = \sum_{h}^{L_A} M_{Ah}\,t_{2h}$ and $\overline{x}_B$ is the ratio mean of all observations for Survey B. Thus, since the $M_{Bh}$ are unknown, there is no explicit stratification of the Survey B frame, i.e., $L_B = 1$.

The separate form considered for L superstrata is

$$\hat{X}_{US} = \sum_{h}^{L} M_{Ah}(t_{1h}\,\overline{x}_{1h} + \theta_h\,t_{2h}\,\overline{x}_{2h} + (1-\theta_h)\,t_{2h}\,\overline{x}_{Bh}) \tag{5.6}$$

The following additional notation is required:

$$\overline{X}_{\ell h} = \sum_{p}^{N_{Ah}} M_{\ell hp}\, \overline{X}_{\ell hp}/M_{\ell h}, \quad \ell = A,1,2,$$

$$\overline{X}_{Bh} = \sum_{p}^{N_{Bh}} M_{Bhp}\, \overline{X}_{Bhp}/M_{Bh},$$

$$\overline{M}_{\ell h} = M_{\ell h}/N_{Ah}, \quad \ell = A,1,2,$$

$$S^2_{\ell h} = \sum_{p}^{N_{Ah}} \frac{M_{\ell hp}}{\overline{M}^2_{\ell h}} (\overline{X}_{\ell hp} - \overline{X}_{\ell h})^2/(N_{Ah} - 1), \quad \ell = A,1,2,$$

$$\tilde{S}^2_{Bh} = \sum_{p}^{N_{Bh}} M_{Bhp}(\overline{X}_{Bhp} - \overline{X}_{Bh})^2/M_{Bh},$$

$$\delta\sigma^2_{\ell h} = \sum_{p}^{N_{Ah}} M_{\ell hp}\, \delta_{\ell hp}\, \sigma^2_{\ell hp}/M_{\ell h}, \quad \ell = A,1,2,$$

$$\delta\sigma^2_{Bh} = \sum_{p}^{N_{Bh}} M_{Bhp}\, \delta_{Bhp}\, \sigma^2_{Bhp}/M_{Bh},$$

$$\tilde{S}^2_{\tau h} = \sum_{p}^{N_{Ah}} M_{Ahp}(\tau_{1hp} - \tau_{1h})^2/M_{Ah}.$$

## 5.1 MSE($\hat{X}_{KC}$) and MSE($\hat{X}_{KS}$)

It is quite common in the literature of survey design optimization to assume simple random sampling without replacement (SRSWOR) of primaries in order to provide a useful form of the between primary variance (see, for example, Hansen, Hurwitz and Madow 1953 and Cochran 1977). If primaries are sampled with unequal probabilities, this approximation tends to overstate the between component. As an alternative, useful approximations can be obtained assuming primaries are sampled with probabilities proportional to size and with replacement (PPSWR). Since this approximation ignores finite population corrections, the between variance may again be overstated in strata where primary sampling fractions are large.

Our approach will be to assume SRSWOR of primaries for the area survey (Survey A) as is customary in the literature, since area strata tend to have small numbers of primaries. However, PPSWR of primaries will be assumed for the telephone survey (Survey B) since, for most designs, the primary finite population corrections are usually negligible. Let $f_{Ah} = n_{Ah}/N_{Ah}$ denote the primary stage sampling fraction. The between component for the separate form (for the combined form, replace $\theta_h$ by $\theta$) is

$$V_1[\sum_p w_{Ahp} \overline{U}_{hp}(\theta_h)] \doteq M_{Ah}^2(1 - f_{Ah})n_{Ah}^{-1} [\theta_h S_{Ah}^2 + (1-\theta_h) \tau_{1h}^2 S_{1h}^2 - \theta_h(1-\theta_h) \tau_{2h}^2 S_{2h}^2]$$

(see Cochran, 1977, p. 250). For Survey B, the between component is

$$V_1(\sum_p w_{Bhp} \overline{X}_{Bhp}) \doteq M_{Bh}^2 n_{Bh}^{-1} \tilde{S}_{Bh}^2$$

The bias in $\hat{X}_{KC}$ and $\hat{X}_{KS}$ is still given by $B(\hat{X}_C)$ and $B(\hat{X}_S)$, respectively. Further, it can be easily verified that variance of an estimator with $\tau_{2h}$ known can be obtained directly from the previous formulas by setting $V(t_{2hp}) = 0$. For simplicity, we give the forms of the variances for the case where $\rho_{12hp} \doteq 0$ (as it is for the special case of simple random sampling within primaries) and approximate $\gamma_B$ defined in (4.2.7) by $M_B^2/k_B^2$.

Therefore,

$$MSE(\hat{X}_{KC}) \doteq B^2(X_C) + \sum_h^{L_A} [(1 - f_{Ah}) \frac{v_{Abh}(\theta)}{n_{Ah}} + \frac{v_{Awh}(\theta)}{n_{Ah} \bar{\bar{m}}_{Ah}}]$$

$$+ \sum_h^{L_B} [\frac{v_{Bbh}(\theta)}{n_{Bh}} + \frac{v_{Bwh}(\theta)}{n_{Bh} \bar{\bar{m}}_{Bh}}] \qquad (5.7)$$

where $\bar{\bar{m}}_{Ah} = E_2(m_{Ah}/n_{Ah})$, $\bar{\bar{m}}_{Bh} = E_2(m_{Bh}/n_{Bh})$,

$$v_{Abh}(\theta) = M_{Ah}^2 \; [\theta \; S_{Ah}^2 + (1-\theta) \; \tau_{1h}^2 \; S_{1h}^2 - \theta(1-\theta) \; \tau_{2h}^2 \; S_{2h}^2]$$

$$v_{Awh}(\theta) = M_{Ah}^2 \; \{\tau_{1h} \; \delta\sigma_{1hp}^2 + \theta^2 \; \tau_{2h} \; \delta\sigma_{2h}^2$$

$$+ q_{Ah} \; \sigma_{ah}^2 \; [(\tau_{1h} + \theta \; \tau_{2h})^2 + (1-\theta)^2 \; \tilde{S}_{\tau h}^2]\}$$

$$v_{Bbh}(\theta) = (1-\theta)^2 \; M_{Bh}^2 \; \tilde{S}_{Bh}^2, \text{ and}$$

$$v_{Bwh}(\theta) = (1-\theta)^2 \; [M_{Bh}^2 \;\; \delta\sigma_{Bh}^2 + M_B^2 \; q_B \; \sigma_\beta^2 \; / L_B].$$

Similarly, we have

$$MSE(\hat{X}_{KS}) \doteq B^2(X_S) + \sum_h^{L} [(1-f_{Ah}) \frac{v_{Abh}(\theta_h)}{n_{Ah}} + \frac{v_{Awh}(\theta_h)}{n_{Ah} \bar{\bar{m}}_{Ah}}$$

$$\frac{v_{Bbh}(\theta_h)}{n_{Bh}} + \frac{v_{Bwh}(\theta_h)}{n_{Bh} \bar{\bar{m}}_{Bh}}]$$

where the form of the variance terms are identical to (5.7), replacing $\theta$ by $\theta_h$.

5.2 MSE($\hat{X}_{UC}$) and MSE($\hat{X}_{US}$)

As in the case of known telephone domain sizes, we have $E(\hat{X}_{UC}) - Y = B(\hat{X}_C)$ (with $L_B = 1$) and $E(\hat{X}_{US}) - Y = B(\hat{X}_S)$. Thus, there is no increase in bias as a result of estimating $\tau_{2h}$ and $M_B$.

Applying the general formulae of previous sections, it can be shown that

$$V(t_{2h}) = (1-f_{Ah})\, n_{Ah}^{-1}\, S_{\tau h}^2 + \overline{m}_{Ah}^{-1}\, \delta\sigma_{\tau h}^2$$

where

$$S_{\tau h}^2 = \sum_p^{N_{Ah}} \frac{M_{Ahp}^2}{\overline{M}_{Ah}^2} (\tau_{2hp} - \tau_{2h})^2/(N_{Ah}-1)$$

and

$$\delta\sigma_{\tau h}^2 = \sum_p^{N_{Ah}} \frac{M_{Ahp}}{M_{Ah}} \phi_{Ahp}\, \tau_{1hp}\, \tau_{2hp}.$$

The mean square error of $\hat{X}_C$ can be obtained directly from (4.2.8) with $L_B = 1$ and adding terms for the covariance between $\hat{M}_B$ and $\hat{X}_{UC}$. The result is given by (5.7) with $L_B = 1$ and replacing $v_{Abh}(\theta)$ by

$$v_{\hat{A}bh}(\theta) = v_{Abh}(\theta) + g_{Abh}(\theta)$$

and $v_{Awh}(\theta)$ by

$$v_{\hat{A}wh}(\theta) = v_{Awh}(\theta) + g_{Awh}(\theta)$$

where

$$g_{Abh}(\theta) = M_{Ah}^2\, S_{\tau h}^2\, [(\overline{X}_{1h} - \theta\, \overline{X}_{2h} - (1-\theta)\, \overline{X}_B)^2 - \overline{D}_h^2\,(\theta)],$$

$\overline{D}_h(\theta) = \overline{X}_{1h} - \theta\, \overline{X}_{2h}$ for the variable $D_{hp}(\theta) = \overline{X}_{1hp} - \theta\, \overline{X}_{2hp}$, and

$$g_{Awh}(\theta) = M_{Ah}^2 \{\sum_p^{M_{Ah}} \frac{M_{Ahp}}{M_{Ah}} \phi_{Ahp}\ \tau_{1hp}\ \tau_{2hp}(\overline{X}_{1hp} - \theta\ \overline{X}_{2hp})^2$$

$$+ \delta\sigma_{\tau h}^2(1-\theta)\ \overline{X}_B\ [(1-\theta)\ \overline{X}_B - 2(\overline{X}_{1h} - \theta\ \overline{X}_{2h})]\}$$

A useful approximation to this form results by assuming $\phi_{Ahp}\ \tau_{1hp}\ \tau_{2hp} \doteq \phi_{Ah}\ \tau_{1h}\ \tau_{2h}$, for average design effect $\phi_{Ah}$. Then $g_{Awh}(\theta)$ becomes

$$g_{Awh}(\theta) = M_{Ah}^2\ \phi_{Ah}\ \tau_{1h}\ \tau_{2h}\ [(\overline{X}_{1h} - \theta\ \overline{X}_{2h} - (1-\theta)\overline{X}_B)^2 + \tilde{S}_{Dh}^2(\theta)] \qquad (5.8)$$

where

$$\tilde{S}_{Dh}^2(\theta) = \sum_p M_{Ahp}\ [\overline{D}_{hp}(\theta) - \overline{D}_h(\theta)]^2/M_{Ah}.$$

Similarly, $MSE(\hat{X}_{US})$ is approximately given by (5.8) replacing $v_{Abh}(\theta_h)$ by

$$\dot{v}_{Abh}(\theta_h) = v_{Abh}(\theta_h) + \dot{g}_{Abh}(\theta_h)$$

and

$v_{Awh}(\theta)$ by

$$\dot{v}_{Awh}(\theta_h) = v_{Awh}(\theta_h) + \dot{g}_{Awh}(\theta_h)$$

where

$$\dot{g}_{Abh}(\theta_h) = M_{Ah}^2\ S_{\tau h}^2[(\overline{X}_{1h} - \theta_h\ \overline{X}_{2h} - (1-\theta_h)\overline{X}_{Bh})^2 - \overline{D}_h^2(\theta_h)]$$

and, in its simplest form,

$$\dot{g}_{Awh}(\theta_h) = M_{Ah}^2\ \phi_{Ah}\ \tau_{1h}\ \tau_{2h}\ [(\overline{X}_{1h} - \theta_h\ \overline{X}_{2h} - (1-\theta_h)\overline{X}_{Bh})^2$$

$$+ \tilde{S}_{Dh}^2(\theta_h)]$$

# 6. MINIMUM MEAN SQUARE ALLOCATION

The procedure for determining the optimum dual frame design is similar in approach to Hartley (1974) applied to a total survey error model and with provisions for dual frame weight, $\theta$, which may vary from stratum to stratum. The optimization method for the separate estimator $\hat{X}_S$, will be discussed in detail. For the combined estimator, changes to Hartley's procedure which allow MSE minimization will be discussed. Where possible, we follow the notation of Hartley.

## 6.1 Optimization with the Separate Estimator

For the general separate estimator in stratified multistage sampling, denote the variance of $\hat{X}_S$

$$V(\hat{X}_S) = \sum_h [V_{Ah}(\underset{\sim}{\alpha}_h, \theta_h) + V_{Bh}(\underset{\sim}{\beta}_h, \theta_h)]$$

where $\underset{\sim}{\alpha}_h$ and $\underset{\sim}{\beta}_h$ are the design vectors for Survey A and Survey B in stratum h. The bias in (4.3.2) does not depend upon $\underset{\sim}{\alpha}_h$ or $\underset{\sim}{\beta}_h$ and is of the form

$$B(\hat{X}_S) = \sum_h B_h(\theta_h)$$

where $B_h(\theta_h)$ is a linear function of $\theta_h$. Hence, the mean square error has the form

$$MSE(\hat{X}_S) = B^2(\hat{X}_S) + V(\hat{X}_S).$$

Denote by $E_{Ah}(\underset{\sim}{\alpha}_h)$ and $E_{Bh}(\underset{\sim}{\beta}_h)$ the expected costs of Survey A and Survey B for stratum h. We wish to determine $\underset{\sim}{\alpha}_h$, $\underset{\sim}{\beta}_h$ and $\theta_h$ for each h so that $MSE(\hat{X}_S)$ is minimized subject to a fixed budget C. This minimization problem will be solved in four stages:

Stage 1: For each stratum h and for given $\theta_h$, $\gamma_h$ and $C_h$

$$\underset{\underset{\sim}{\alpha}_h}{\text{minimize}} \qquad V_{Ah}(\underset{\sim}{\alpha}_h, \theta_h)$$

$$\text{subject to} \qquad E_A(\underset{\sim}{\alpha}_h) = \gamma_h C_h$$

and

$$\underset{\underset{\sim}{\beta}_h}{\text{minimize}} \qquad V_{Bh}(\underset{\sim}{\beta}_h, \theta_h)$$

$$\text{subject to} \qquad E_B(\underset{\sim}{\beta}_h) = (1 - \gamma_h) C_h$$

where $C_h$ is a given budget to be allocated to stratum h and $\gamma_h$ is a given fraction of $C_h$ to be allocated to Survey A, $0 < \gamma_h < 1$ and hence, $1 - \gamma_h$ is the fraction to be allocated to Survey B in stratum h.

The solutions to these mathematical programs yield

$$v_{Ah}(C_h, \gamma_h, \theta_h) \text{ and } v_{Bh}(C_h, \gamma_h, \theta_h),$$

the minimum conditional variances in each stratum.

Stage 2: For given $C_h$ and $\theta_h$, perform the following minimization:

$$\underset{\gamma_h}{\text{minimize}} \qquad \{v_{Ah}(C_h, \gamma_h, \theta_h) + v_{Bh}(C_h, \gamma_h, \theta_h)\}$$

which will yield conditional variance $v_h(C_h, \theta_h)$ to be entered in Stage 3.

Stage 3: For given $\theta_h$,

$$\underset{C_h}{\text{minimize}} \qquad \sum_h v_h(C_h, \theta_h).$$

$$\text{subject to} \qquad \sum_h C_h = C.$$

We denote this conditional minimum by $V(\underset{\sim}{\theta})$ where $\underset{\sim}{\theta} = [\theta_1, \ldots, \theta_L]$.

Finally, we solve

Stage 4:

$$\underset{\underset{\sim}{\theta}}{\text{minimize}} \qquad M\,(\underset{\sim}{\theta}) \tag{6.1.1}$$

where $M\,(\underset{\sim}{\theta}) = [\Sigma\; B_h\;(\theta_h)]^2 \;+\; V(\underset{\sim}{\theta})$.

In Appendix B, it is established that the procedure yields the global minimum of $MSE(\hat{X}_S)$. The following application should clarify the procedure.

Consider a stratified two-stage telephone/area dual frame survey with MSE given by any of the forms in Section 5. If we ignore all fpc's, then

$$V_{Ah}\;(\underset{\sim}{\alpha}_h,\; \theta_h) \;=\; \frac{v_{Abh}(\theta_h)}{\alpha_{1h}} + \frac{v_{Awh}\;(\theta_h)}{\alpha_{1h}\alpha_{2h}} \tag{6.1.2}$$

and

$$V_{Bh}\;(\underset{\sim}{\beta}_h,\; \theta_h) \;=\; \frac{v_{Bbh}\;(\theta_h)}{\beta_{1h}} + \frac{v_{Bwh}\;(\theta_h)}{\beta_{1h}\;\beta_{2h}} \tag{6.1.3}$$

where $\underset{\sim}{\alpha}_h \;=\; [n_{Ah},\; \bar{\bar{m}}_{Ah}]$ and $\underset{\sim}{\beta}_h = [n_{Bh},\; \bar{\bar{m}}_{Bh}]$. Let the cost function be of the form,

$$C = \sum_h (C_{Abh}\alpha_{1h} + C_{Awh}\alpha_{1h}\alpha_{2h} + C_{Bbh}\beta_{1h}$$

$$+\; C_{Bwh}\beta_{1h}\beta_{2h})$$

where $C_{Abh}$, $C_{Bbh}$ are the respective costs associated with primary units and $C_{Awh}$, $C_{Bwh}$ are the respective costs associated with secondary units in each survey.

For Stage 1, we apply the Lagrangian method with the following results:

$$\alpha_{1h} = \gamma_h C_h (v_{Abh}(\theta_h)/C_{Abh})^{\frac{1}{2}}/a(\theta_h),$$

$$\alpha_{2h} = [(v_{Awh}(\theta)/C_{Awh})(C_{Abh}/v_{Abh}(\theta_h))]^{\frac{1}{2}}$$

$$\beta_{1h} = (1 - \gamma_h) C_h (v_{Bbh}(\theta_h)/C_{Bbh})^{\frac{1}{2}}/b(\theta_h),$$

$$\beta_{2h} = [(v_{Bwh}(\theta)/C_{Bwh})(C_{Bbh}/v_{Bbh}(\theta_h))]^{\frac{1}{2}}$$

where

$$a(\theta_h) = (C_{Abh}\, v_{Abh}(\theta_h))^{\frac{1}{2}} + (C_{Awh}\, v_{Awh}(\theta_h))^{\frac{1}{2}}$$

and

$$b(\theta_h) = (C_{Bbh}\, v_{Bbh}(\theta_h))^{\frac{1}{2}} + (C_{Bwh}\, v_{Bwh}(\theta_h))^{\frac{1}{2}}.$$

By substitution into (6.1.2) and (6.1.3), we have

$$v_{Ah}(C_h, \gamma_h, \theta_h) = a^2(\theta_h)/(\gamma_h C_h), \qquad (6.1.4)$$

and

$$v_{Bh}(C_h, \gamma_h, \theta_h) = b^2(\theta_h)/[(1 - \gamma_h)C_h], \qquad (6.1.5)$$

Entering (6.1.4) and (6.1.5) into Stage 2 produces the optimum fraction $\gamma_h$ to be allocated to the area frame survey, viz.

$$\gamma_h = a(\theta_h)\,[a(\theta_h) + b(\theta_h)]^{-1}$$

and substitution into the sum of (6.1.4) and (6.1.5) yields

$$v_h(C_h, \theta_h) = (a(\theta_h) + b(\theta_h))^2/C_h \qquad (6.1.6)$$

For Stage 3, the application of the Lagrangian method yields

$$C_h = C\,[a(\theta_h) + b(\theta_h)]\ \{\sum_h [a(\theta_h) + b(\theta_h)]\}^{-1}$$

with conditional variance from (6.1.6), summing over all strata, given by

$$V(\underset{\sim}{\theta}) = \{\sum_h [a(\theta_h) + b(\theta_h)]\}^2/C.$$

which is similar in form to Hartley's equation 12.

For the final stage, we minimize

$$M(\underset{\sim}{\theta}) = [\sum_h B_h(\theta_h)]^2 + V(\underset{\sim}{\theta})$$

where $B_h(\theta_h) = E(X_S) - Y$ given by (4.3.2).

An analytic minimization of $M(\underset{\sim}{\theta})$ is not feasible, however, any convex programming algorithm will yield the optimal dual frame weights $\theta_1^*, \ldots, \theta_L^*$.

## 6.2 Optimization with the Combined Estimator.

The problem of optimal dual frame survey allocation for the general combined estimator was treated in Hartley (1974) for the case of no nonsampling errors and unbiased estimators. When nonsampling errors are introduced, his procedure for optimization remains the same for the first two stages. For these stages, bias component is ignored. However, in the third stage, the conditional mean square error $M(\theta)$ is minimized instead of $v(\theta)$ in his equation 12, where

$$M(\theta) = B(\theta) + v(\theta)$$

and $B(\theta) = E(X_C) - Y$ is given by (4.2.3).

## 7. AN EXAMPLE

The Current Population Survey (CPS) is a household sample survey conducted monthly by the Bureau of the Census to provide estimates of unemployment and other characteristics of the general labor force. To illustrate the minimum mean square optimization method of Section 6, the required cost and error parameters were estimated for CPS assuming 51 superstrata corresponding to the 50 states and the District of Columbia.

The CPS sample design is essentially a stratified two-stage design with counties as primary sampling units and areal segments of dwelling units as secondaries. The following estimators of the U.S. monthly unemployment rate and their approximate MSE formulae obtained from Section 5 shall be illustrated and compared:

1. <u>Separate Estimator, $\tau_{2h}$ Known</u>

$$\bar{x}_{KS} = \sum_h W_h(\tau_{1h}\,\bar{x}_{1h} + \theta_h\,\tau_{2h}\,\bar{x}_{2h} + (1-\theta_h)\,\tau_{2h}\,\bar{x}_{Bh})$$

and

$$\mathrm{MSE}(\bar{x}_{KS}) \doteq [\sum_h W_h(b_{Ah}(\theta_h) + \theta_h\,B_{Bh})]^2$$

$$+ \sum_h W_h^2[n_{Ah}^{-1}\,v_{Ah}(\theta_h) + n_{Bh}^{-1}\,v_{Bh}(\theta_h)]$$

where $W_h = M_{Ah}/\sum_h M_{Ah}$, $b_{Ah}(\theta_h)$ and $B_{Bh}$ are as defined in Section 5.2,

$$v_{Ah}(\theta_h) = \theta_h\,S_{Ah}^2 + (1-\theta_h)\,\tau_{1h}^2\,S_{1h}^2 - \theta_h(1-\theta_h)\,\tau_{2h}^2\,S_{2h}^2$$

$$+ \bar{\bar{m}}_{Ah}^{-1}\,\{\tau_{1h}\,\delta_{1h}\,\sigma_{1h}^2 + \theta_h^2\,\tau_{2h}\,\delta_{2h}\,\sigma_{2h}^2$$

$$+ [(\tau_{1h} + \theta_h \tau_{2h})^2 + (1-\theta_h)^2 (\Phi_{Ah} - \phi_{Ah}) \tau_{1h} \tau_{2h}] \sigma^2_{\alpha h} q_{Ah}\},$$

and

$$\bar{\bar{m}}_{Bh} v_{Bh}(\theta_h) = (1-\theta_h)^2 (\tau^2_{2h} \delta_{Bh} \sigma^2_{Bh} + \tau^2_2 \sigma^2_{\beta h} q_B/L).$$

In the above formulas, $\delta_{1h}$, $\delta_{2h}$, $\delta_{Bh}$ are average with primary design effects, $\sigma^2_{1h}$, $\sigma^2_{2h}$, $\sigma^2_{Bh}$ are average within primary sample random sampling variances, $\Phi_{Ah}$ is the total sample design effect and $\phi_{Ah}$ is the average within primary design effect associated with $t_{1h}$ defined in (5.8).

2. Separate Estimator, $\tau_{2h}$ Unknown

$$\bar{x}_{US} = \sum_h W_h(t_{1h} \bar{x}_{1h} + \theta_h t_{2h} \bar{x}_{2h} + (1-\theta_h) t_{2h} \bar{x}_{Bh})$$

and $MSE(\bar{x}_{US}) \doteq MSE(\bar{x}_{KS})$ replacing $v_{Ah}(\theta_h)$ by

$$\dot{v}_{Ah}(\theta_h) = v_{Ah}(\theta_h) + \bar{\bar{m}}^{-1}_{Ah} \dot{g}_{Ah}(\theta_h)$$

where

$$\dot{g}_{Ah}(\theta_h) = \tau_{1h} \tau_{2h} \{\Phi_{Ah}[(\bar{X}_{1h} - \theta_h \bar{X}_{2h} - (1-\theta_h) \bar{X}_{Bh})^2$$

$$- (\Phi_{Ah} - \phi_{Ah}) (\bar{X}_{1h} - \theta_h \bar{X}_{2h})^2]$$

3. Combined Estimator, $\tau_{2h}$ Known

$$\bar{x}_{KC} = \sum_h W_h(\tau_{1h} \bar{x}_{1h} + \theta \tau_{2h} \bar{x}_{2h}) + (1-\theta) \tau_2 \bar{x}_B$$

and

$$MSE(\bar{x}_{KC}) \doteq [\sum_h W_h b_{Ah}(\theta) + \theta B_B]^2 + \sum_h W^2_h n^{-1}_{Ah} v_{Ah}(\theta) + n^{-1}_B v_B(\theta)$$

where $b_{Ah}(\theta)$ and $B_B(\theta)$ are as defined in (4.2.3) for $L_B = 1$, $v_{Ah}(\theta) = v_{Ah}(\theta_h)$ with $\theta_h = \theta$ for all h, and

$$\bar{\bar{m}}_B \, v_B(\theta) = (1-\theta)^2 \, \tau_2 \, (\delta_B \, \sigma_B^2 + q_B \, \sigma_\beta^2)$$

where $\delta_B$ is the total sample design effect and $\sigma_B^2$ is the sample random sampling variance for Survey B.

4. Combined Estimator, $\tau_{2h}$ Unknown

$$\bar{x}_{UC} = \sum_h W_h \, (t_{1h} \, \bar{x}_{1h} + \theta \, t_{2h} \, \bar{x}_{2h}) + (1-\theta) \, t_2 \, \bar{x}_B$$

and $MSE(\bar{x}_{UC}) \doteq MSE(\bar{x}_{KC})$ replacing $v_{Ah}(\theta)$ with

$$v_{\hat{A}h}(\theta) = v_{Ah}(\theta) + \bar{\bar{m}}_{Ah}^{-1} \, g_{\hat{A}h}(\theta)$$

where

$$g_{\hat{A}h}(\theta) = \tau_{1h} \, \tau_{2h} \, \{\Phi_{Ah}[(\bar{X}_{1h} - \theta \, \bar{X}_{2h} - (1-\theta) \, \bar{X}_B)^2$$

$$- (\Phi_{Ah} - \phi_{Ah}) \, (\bar{X}_{1h} - \theta \, \bar{X}_{2h})^2]$$

7.1 Data Set

In 1982, the average CPS interviewer workload was about 50 dwellings and currently the average number of interviewer assignments per PSU (primary sampling unit) is about 2. These parameters, which determine the $q_{Ah}$ and $\bar{\bar{m}}_{Ah}$, shall be held fixed. Thus, for Survey A, only $n_{Ah}$, the number of PSUs per stratum, is to be optimized.

In current RDD studies at the Census Bureau, the Waksberg (1978) sampling method is used with the within primary sampling quota equal to six residential telephone numbers. This parameter, which determines the $\bar{\bar{m}}_{Bh}$, shall also be held fixed. Then only $n_{Bh}$(or $n_B$) need be optimized for Survey B.

The following simple cost models shall be used in the optimization:

$$C = C_o + \sum_h [100 \, C_{Ah} \, n_{Ah} + 6 \, C_{Bh} \, n_{Bh}]$$

for the separate estimator and

$$C = C_o + \sum_h 100 C_{Ah} n_{Ah} + 6 C_B n_B$$

for the combined estimator where C is the total survey budget, $C_o$ is the fixed survey cost and $C_{Ah}$ and $C_{Bh}$ ($C_B$) are the average variable cost per unit for Survey A and Survey B interviewers, respectively. For simplicity we shall assume that the total fixed costs, $C_o$, will not change over current levels for a dual frame survey. Thus, we only need to know $C_{Ah}$, $C_{Bh}$ ($C_B$), and $VC = C - C_o$ in order to optimize $n_{Ah}$ and $n_B$.

The population target parameter is the CPS monthly unemployment rate. For nontelephone and telephone households, the assumed rates are $P_1 = 15\%$ and $P_2 = 6\%$, respectively. (This relative difference is consistent with available data.) We shall assume they are the same in all strata, as are the within primary design effects $\delta_{Ah} = 1.33$ and $\delta_{Bh} = 1.25$, which were computed from available data for the CPS. We use the usual sample sampling formulas to compute $\sigma_1^2$ and $\sigma_2^2$ and the domain design effects $\delta_{1h}$ and $\delta_{2h}$ are obtained using the formulas in Appendix A.

Table 1 summarizes the optimization parameters that vary across strata. The telephone coverage rates, $\tau_{2h}$, are proportions of households with telephones from the 1980 Census. The between PSU variance components, $S_{Ah}^2$, are obtained from CPS as the percent of $S_{Ah}^2$ to $S_{Ah}^2 + \delta_{Ah} \bar{\bar{\sigma}}_{Ah}^2/m_A$, denoted by $BPSU_h$. We shall assume that $S_{1h}^2 \doteq S_{2h}^2 \doteq S_{Ah}^2$.

Variable costs per household by state ($C_{Ah}$) were synthetically estimated for Survey A using available data on regional costs per household, interviewer time and mileage by state, and CPS state workloads. The stratum weights, $W_h$ are based on recent CPS data on civilian labor force by state.

There is little information available on biases for area/list or telephone frame surveys. To simulate the effects of biases which may vary by state for each frame, we shall assume that biases are proportional to Survey A nonresponse rates ($NR_h$).

The parameter values in the optimization are summarized in Tables 1 and 2.

(Insert tables 1 and 2 about here.)

7.2 Optimization Results

The results of the dual frame survey optimization procedure for the separate estimator are given in the last two columns of Table 2. The MSE of $\overline{x}_{US}$ was minimized subject to total variable costs, VC, set at about \$1 million which is $\frac{1}{12}$ the estimated total annual variable cost for CPS. $ALLOC_h$, the optimal allocation of sample to the telephone survey, is reported in the next-to-last column for the case of zero nonsampling biases. For example, for estimating monthly unemployment, seven states would not use the telephone frame ($ALLOC_h = 0$) while 27 states would allocate at least 50% to RDD. The total telephone sample allocation is 23%.

Suppose a small differential bias between the two surveys, say 5% of the proportion to be estimated, is assumed. In the last column, TBIAS, the telephone survey bias parameter, is 5% while ABIAS, the area survey bias parameter, 0. Now $ALLOC_h = 0$ for 29 states and only 1 state would allocate as much as 50% to the telephone survey. Nationally, only 3% of the sample would be allocated to RDD. The table illustrates how widely $ALLOC_h$ can vary between states as well as the potential impact of telephone bias and other survey parameters on dual frame allocation.

Using the data in Table 1 and 2, in Table 3 we compare the relative efficiency of the four dual frame estimators. Here, our measure of efficiency is the reduction in MSE of the estimator from the minimum MSE for the single frame design. Even though the reduction is small for all cases, it is at least twice as great for the separate estimators than for the combined estimators for both values of TBIAS. It also appears that, for these data, the effect of estimating $\tau_{2h}$ in each stratum by $t_{2h}$ (for $\overline{x}_{US}$ and $\overline{x}_{UC}$) is a relatively small loss in efficiency over the case of $\tau_{2h}$ known ($\overline{x}_{KS}$ and $\overline{x}_{KC}$).

(Insert table 3 about here.)

## Appendix A - Form of the Within Primary Design Effects

We consider $\delta_{1hp}$, the within primary design effect for $D_1$ elements in Survey A. The forms of $\delta_{2hp}$ and $\delta_{Bhp}$ can be obtained analogously.

Let

$$y'_{1hpst} = y_{1hpst} + \epsilon_{1hpst}$$

and define the following notation:

$j_{hp}$ = number of secondaries samples in primary (h,p),

$J_{hp}$ = total number of secondaries in primary (h,p),

$M_{1hps}$ = number of elements in $D_1$ for the s-th secondary in primary (h,p),

$m_{1hp}$ = $\sum_{s}^{j_{hp}} Q_{1hps}$, the total number of $D_1$ elements in the sample for primary (h,p),

$\bar{\bar{m}}_{1hp}$ = $E(m_{1hp}|h,p)$, the expected number of sample elements in $D_1$ for primary (h,p),

$M_{1hp}$ = total number of $D_1$ elements in primary (h,p),

$y'_{1hp}$ = $\sum_{s}^{j_{hp}} \sum_{t}^{Q_{1hps}} y_{1hpst}$, sample total for the primary,

$\bar{y}'_{1hp}$ = $y'_{1hp}/m_{1hp}$, sample mean per element for the primary, and

$\tilde{y}'_{1hp}$ = $y'_{1hp}/j_{hp}$, sample mean per secondary for the primary,

$\tilde{m}_{1hp}$ = $m_{1hp}/j_{hp}$, average number of sample $D_1$ elements per secondary,

$\bar{Y}'_{1hp}$ = $E(y'_{1hpst}|h,p)$.

The large sample (Taylor's series) approximation of $V(\bar{y}'_{1hp})$ is

$$\frac{1}{E^2(\tilde{m}_{1hp})} V(\tilde{y}'_{1hp} - \bar{Y}'_{1hp}\,\tilde{m}_{1hp})$$

$$= \frac{1}{E^2(\tilde{m}_{1hp})} V(\tilde{\Delta}_{1hp}) \tag{A.1}$$

where $\Delta_{1hps} = \sum_{t \in D_1} (y'_{1hpst} - \bar{Y}'_{1hp})$ and $\tilde{\Delta}_{1hp} = \sum_s \Delta_{1hps}/j_{hp}$.

Consider the special case where secondary stage units are drawn with SRSWR. Then $E(\tilde{m}_{1hp}) = M_{1hp}/J_{hp}$ and $\bar{\bar{m}}_{1hp} = j_{hp}\, M_{1hp}/J_{hp}$. (A.1) now has the form

$$\frac{J^2_{hp}}{M^2_{1hp}} \frac{1}{j_{hp}} E(\Delta^2_{1hps})$$

$$= \frac{1}{\bar{\bar{m}}_{1hp}} \sigma^2_{1hp} \left(1 + \frac{1}{m_{1hp}} g_{1hp}\, \lambda_{1hp}\right)$$

where

$$\lambda_{1hp}\, \sigma^2_{1hp} = \frac{\sum_s^{J_{hp}} \sum_t \sum_{t \neq t'} (y_{1hpst} - \bar{Y}_{1hp})(y_{1hpst'} - \bar{Y}_{1hp})}{g_{1hp}},$$

$$g_{1hp} = \sum_s^{J_{hp}} M_{1hps}\,(M_{1hps} - 1),$$

and $\sigma^2_{1hp}$ is as defined in Section 4.2. Note that $g_{1hp}/M_{1hp}$ may also be written as

$$m_{1hp}\ \mathrm{Relvar}(\tilde{m}_{1hp}) + E(\tilde{m}_{1hp})-1$$

Therefore, the design effect $\delta_{1hp}$ defined in (4.2.5) has the form

$$\delta_{1hp} = 1 + \lambda_{1hp}[(\mathrm{Relvar}(M_{1hps}) + 1)\ E(\tilde{m}_{1hp})-1]$$

which takes the familiar form

$$\delta_{1hp} = 1 + \lambda_{1hp}[E(\tilde{m}_{1hp})-1] \qquad (A.2)$$

when the relative variance of the $M_{1hps}$ is negligible. This same procedure can be applied to $\delta_{1hp}$ and $\delta_{Bhp}$. Note that, for the particular application in this paper, $\lambda_{1hp}$ and $\lambda_{2hp}$ are intra-segment correlation coefficients while $\lambda_{Bhp}$ is a intra-household correlation.

## Appendix B - Proof of Optimality

We shall establish that under very general conditions the four stage optimization procedure of Section 6 produces the global minimum of the constrained function

$$M(\underset{\sim}{\eta}, \underset{\sim}{\theta}) = \sum_h [V_{Ah} (\underset{\sim}{\alpha}_h, \theta_h) + V_{Bh} (\underset{\sim}{\beta}_h, \theta_h)]$$

$$+ [\sum_h B_h (\theta_h)]^2 \qquad (B.1)$$

subject to

$$C(\underset{\sim}{\eta}) = \sum_h [E_{Ah} (\underset{\sim}{\alpha}_h) + E_{Bh} (\underset{\sim}{\beta}_h)] = C$$

where $\underset{\sim}{\eta} = [\underset{\sim}{\alpha}_h, \underset{\sim}{\beta}_h]$ and $\underset{\sim}{\theta}$ are the variate vectors and $C(\underset{\sim}{\eta})$ is a linear function of $\underset{\sim}{\eta}$. Clearly, the method produces a stationary point of the function (B.1). To establish that this point is the unique minimum point, it suffices to prove (1) the strict convexity of (B.1) for general survey designs and (2) the strict convexity of $M(\underset{\sim}{\theta})$ (defined in (5.1.1)) as a function of $\underset{\sim}{\theta}$.

<u>Theorem 1</u>: The variance functions

$$V_{Ah} (\alpha_h, \theta_h) = V_1 \sum_p w_{Ahp} \overline{U}_{hp} (\theta_h) + E_1 [V_{AWh}(\theta_h; \Pi_A)]$$

and

$$V_{Bh} (\alpha_h, \theta_h) = (1 - \theta_h)^2 \{V_1 \sum_p W_{Bhp} \overline{X}_{Bhp} + E_1 [V_{Bwh}(\Pi_B)]\}$$

given in (4.3.3) are strictly convex functions when sampling fractions are ignored at all stages.

<u>Proof</u>: Consider the between primary component of $V_{Ah}(\alpha_h; \theta_h)$. Ignoring sampling fractions, sampling theory without nonsampling error provides formulas of the form

$$V\left(\sum_p w_{Ahp}\,\overline{U}_{hp}\right) = \sum_{t=1}^{k} \alpha_t^{-1}\,\overline{\underset{\sim}{U}}_h'\,\underset{\sim}{Q}_t\,\overline{\underset{\sim}{U}}_h$$

where $\overline{\underset{\sim}{U}}_h$ is the vector of primary population means, the $Q_t$ are symmétrical and non-degenerate positive definite matrices, k + 1 is the number of sampling stages, and $\alpha_t$ is the number of units drawn at the tth stage above the last stage. Thus, applying Hartley's (1974) Theorem 1 proves $V\left(\Sigma\, w_{Ahp}\,\overline{U}_{hp}\right)$ is strictly convex.

Consider now the within primary component of $V_{Ah}(\underset{\sim}{\alpha}_h, \theta_h)$. This term may be written as

$$E_1 \sum_p \alpha_p^{-1}\, w_{Ahp}^2 \left(f_{1hp} + \theta_h^2\, f_{2hp} + 2\theta_h\, f_{12hp}\right) \qquad \text{(B.2)}$$

where

$$f_{1hp} = \tau_{1hp}\,\delta_{1hp}\,\sigma_{1hp}^2 + \phi_{Ahp}\,\tau_{1hp}\,\tau_{2hp}\,\overline{X}_{1hp}^2 + \tau_{1hp}^2\,\sigma_{\alpha h}^2\,q_{Ah}$$

$$f_{2hp} = \tau_{2hp}\,\delta_{2hp}\,\sigma_{2hp}^2 + \phi_{Ahp}\,\tau_{1hp}\,\tau_{2hp}\,\overline{X}_{2hp}^2 + \tau_{2hp}^2\,\sigma_{\alpha h}^2\,q_{Ah}$$

and

$$f_{12hp} = \left(\tau_{1hp}\,\tau_{2hp}\,\delta_{1hp}\,\delta_{2hp}\right)^{1/2}\sigma_{1hp}\,\sigma_{2hp}\,\rho_{12hp} - \phi_{Ahp}\,\tau_{1hp}\,\tau_{2hp}\,\overline{X}_{1hp}\,\overline{X}_{2hp} + \tau_{1hp}\,\tau_{2hp}\,\sigma_{\alpha h}^2\,q_{Ah}$$

The $(n_{Ah} + 1) \times (n_{Ah} + 1)$ Hessian matrix for the summation in (B.2) before expectation has the form

$$\underset{\sim}{H} = 2w_{Ahp}^{2} \begin{bmatrix} g_{11} & -\underset{\sim}{g}' \\ -\underset{\sim}{g} & \underset{\sim}{D} \end{bmatrix}$$

where $$g_{11} = \sum_{p} \alpha_p^{-1} \, f_{2hp},$$

$$\underset{\sim}{g}' = [g_p] \text{ with, for } p = 1,\ldots,n_{Ah},$$

$$g_p = \alpha_p^{-2} \, (\theta_h \, f_{2hp} + f_{12hp}),$$

and $$\underset{\sim}{D} = \text{diag } (d_p) \text{ with, for } p = 1,\ldots,n_{Ah},$$

$$d_p = \alpha_p^{-3} \, (f_{1hp} + \theta_{hp}^{2} \, f_{2hp} + 2\theta_h \, f_{12hp}).$$

Clearly, since all $d_p > 0$, the principal minor determinants of order $n_{Ah}$ or less are positive. Therefore, we must show $\det(\underset{\sim}{H}) > 0$ in order to establish that H is positive definite.

Using the well-known identity

$$\det(\underset{\sim}{H}) = \det(\underset{\sim}{D})(g_{11} - \underset{\sim}{g}' \, \underset{\sim}{D}^{-1} \, \underset{\sim}{g}),$$

it can be shown that $\det(\underset{\sim}{H}) > 0$ if

$$f_{1hp} \, f_{2hp} - f_{12hp}^{2} =$$

$$\tau_{1hp} \, \tau_{2hp} \, \delta_{1hp} \, \delta_{2hp} \, \sigma_{1hp} \, \sigma_{2hp} \, (1 - \rho_{12hp}^{2})$$

$$+ \; \tau_{1hp}^{2} \, \tau_{2hp}^{2} \, \sigma_{\alpha h}^{2} \, q_{Ah} \left[\frac{\delta_{1hp} \, \sigma_{1hp}^{2}}{\tau_{1hp}} + \frac{\delta_{2hp} \, \sigma_{2hp}^{2}}{\tau_{2hp}}\right]$$

$$- 2 \left(\frac{\delta_{1hp}\ \delta_{2hp}}{\tau_{1hp}\ \tau_{2hp}}\right)^{1/2} \sigma_{1hp}\ \sigma_{2hp}\ \rho_{12hp})]$$

$$+ \text{(positive cross product terms)} > 0.$$

The first term on the right is positive since $\rho_{12hp} < 1$ and the second term is positive since the term in parentheses is equivalent to $\overline{m}_{Ahp}$ times $V_2(\overline{x}_{1hp}) + V_2(\overline{x}_{2hp}) - 2\ Cov_2(\overline{x}_{1hp}, \overline{x}_{2hp}) = V_2(\overline{x}_{1hp} - \overline{x}_{2hp}) > 0$. This proves the strict convexity of (B.2). Since the sum of convex functions is convex, the expected value, $V_w$, is convex and $V_{Ah}\ (\underset{\sim}{\alpha}_h, \theta_h)$ is a strictly convex function. A similar argument applied to $V_{Bh}\ (\underset{\sim}{\beta}_h, \theta_h)$ proves the theorem.

<u>Theorem 2</u>: Under the conditions of Theorem 1, the constrained function

$$M_C\ (\underset{\sim}{\eta}, \underset{\sim}{\theta}) = \{M\ (\underset{\sim}{\eta}, \underset{\sim}{\theta}) | C(\underset{\sim}{\eta}) = C\},$$

for linear cost functions $C(\underset{\sim}{\eta})$, is a strictly convex function of $\underset{\sim}{\eta}$ and $\underset{\sim}{\theta}$.

<u>Proof</u>: By Theorem 1, $V_{Ah}\ (\underset{\sim}{\alpha}_h, \theta_h)$ and $V_{Bh}\ (\underset{\sim}{\beta}_h, \theta_h)$ are strictly convex variance functions. Hence, the sum of the components over the L strata is strictly convex.

Since $B_h\ (\theta_h)$ are linear functions of $\theta_h$, $(\Sigma\ B_h\ (\theta_h))^2 = \underset{\sim}{B}'J\ \underset{\sim}{B}$ where $\underset{\sim}{B}$ is the L-vector of the functions $B_h\ (\theta_h)$ and J is the L x L matrix of 1's, has a positive definite Héssian matrix with respect to $\underset{\sim}{\theta}$ and is, therefore, convex. The sum of convex functions is convex and, thus, convex over a linear subspace. Thus, $M_C(\underset{\sim}{\eta}, \underset{\sim}{\theta})$ is strictly convex.

<u>Theorem 3</u>: Under the conditions of Theorem 1, the function

$$M(\underset{\sim}{\theta}) = \min_{\underset{\sim}{\eta}|\underset{\sim}{\theta}} M_C(\underset{\sim}{\eta}, \underset{\sim}{\theta})$$

is strictly convex.

Proof: Because of Theorem 1, the program

$$\min_{\underset{\sim}{\eta} \mid \underset{\sim}{\theta}} \sum_h \left[V_{Ah} (\underset{\sim}{\alpha}_h, \theta_h) + V_{Bh} (\underset{\sim}{\beta}_h, \theta_h)\right]$$

$$\text{s.t.} \quad C (\underset{\sim}{\eta}) = C$$

has a unique solution vector which will be denoted as $\underset{\sim}{\eta}^* (\underset{\sim}{\theta})$. Choose two $\underset{\sim}{\theta}$-vectors, say $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\theta}_2$, and two constants $\lambda_1 > 0$ and $\lambda_2 > 0$ with $\lambda_1 + \lambda_2 = 1$. From Theorem 2, $M_C (\underset{\sim}{\eta}, \underset{\sim}{\theta})$ is strictly convex. Hence, we have

$$\begin{aligned} & M_C (\lambda_1 \underset{\sim}{\eta}^* (\underset{\sim}{\theta}_1) + \lambda_2 \underset{\sim}{\eta}^* (\underset{\sim}{\theta}_2), \lambda_1 \underset{\sim}{\theta}_1 + \lambda_2 \underset{\sim}{\theta}_2) \\ & < \lambda_1 M_C (\underset{\sim}{\eta}^* (\underset{\sim}{\theta}_1), \underset{\sim}{\theta}_1) + \lambda_2 M_C (\underset{\sim}{\eta}^* (\underset{\sim}{\theta}_2), \underset{\sim}{\theta}_2) \\ & = \lambda_1 M (\theta_1) + \lambda_2 M (\theta_2). \end{aligned} \tag{B.3}$$

But, $M (\lambda_1 \underset{\sim}{\theta}_1 + \lambda_2 \underset{\sim}{\theta}_2)$ cannot be larger than (B.3) since it is the minimum of $M_C (\underset{\sim}{\eta}, \lambda_1 \underset{\sim}{\theta}_1 + \lambda_2 \underset{\sim}{\theta}_2)$ over all $\underset{\sim}{\eta}$ . This proves $M (\underset{\sim}{\theta})$ is strictly convex.

1. Parameter Summary

| Parameter | Value |
|---|---|
| Total variable costs | $VC = \$997{,}864.00$/month |
| Survey B costs/unit | $C_B = \$11.50$ |
| Survey A deffs | $\delta_{Ah} = 1.33,\ h=1,\ldots,51$ |
| Domain $D_1$ | $\delta_{1h} = 1 + (\tau_{1h}\,\bar{\bar{m}}_{Ah}-1)(\delta_{Ah}-1)/(\bar{\bar{m}}_{Ah}-1)$ |
| Domain $D_2$ | $\delta_{2h} = 1 + (\tau_{2h}\,\bar{\bar{m}}_{Ah}-1)(\delta_{Ah}-1)/(\bar{\bar{m}}_{Ah}-1)$ |
| | $\phi_{Ah} = 1,\ h=1,\ldots,51$ |
| | $\Phi_{Ah} = 1,\ h=1,\ldots,51$ |
| Survey B deff | $\delta_{Bh} = 1.25,\ h=1,\ldots,51$ |
| Proportion with characteristic | $P_{Ah} = \tau_{1h} P_1 + \tau_{2h} P_2$ |
| Domain $D_1$ | $P_1 = .15$ |
| Domain $D_2$ | $P_2 = .06$ |
| Between PSU variance | $S^2_{Ah} = \dfrac{BPSU_h}{1-BPSU_h}\,\delta_{Ah}\,P_{Ah}(1-P_{Ah})/\bar{\bar{m}}_{Ah}$ |
| Domain $D_1$ | $S^2_{1h} = S^2_{Ah}$ |
| Domain $D_2$ | $S^2_{2h} = S^2_{Ah}$ |
| Survey A bias | Let $g_h = NR_h/(\Sigma\, W_h\, NR_h)$, then |
| Domain $D_1$ | $B_{1h} = g_h \cdot ABIAS \cdot P_1$ |
| Domain $D_2$ | $B_{2h} = g_h \quad ABIAS \cdot P_2$ |
| Survey B bias | $B_{Bh} = g_h \quad TBIAS \cdot P_2$ |
| Household size | $HHSIZE = 1.12$ civilian labor force/household |
| Interviewer Assignment Size Survey A | $q_{Ah} = 50 \cdot HHSIZE,\ h=1,\ldots,51$ |
| Survey B | $q_B = 90 \cdot HHSIZE$ |
| Within PSU sample sizes Survey A | $\bar{\bar{m}}_{Ah} = 133 \cdot HHSIZE,\ h=1,\ldots,51$ |
| Survey B | $\bar{\bar{m}}_{Bh} = 6 \cdot HHSIZE,\ h=1,\ldots,51$ |

.2. CPS Data Set and Optimum Telephone Allocation for TBIAS = 0% and 5%.

| Stratum | Parameter Values | | | | | $ALLOC_h$ | |
|---|---|---|---|---|---|---|---|
| | $W_h$ | $\tau_{2h}$ | $BPSU_h$ | $C_{Ah}$ | $NR_h$ | TBIAS=0% | TBIAS=5% |
| US | 100% | 93% | 10% | $14 | 4% | 23% | 3% |
| AL | 1.5 | 87 | 9 | 17 | 3 | 43 | 10 |
| AK | .2 | 84 | 8 | 20 | 6 | 54 | 8 |
| AZ | 1.4 | 90 | 2 | 17 | 5 | 45 | 0 |
| AR | 1.0 | 87 | 9 | 20 | 3 | 51 | 39 |
| CA | 11.0 | 95 | 2 | 9 | 5 | 0 | 0 |
| CO | 1.6 | 94 | 6 | 16 | 4 | 58 | 0 |
| CT | 1.4 | 97 | 0 | 15 | 4 | 60 | 0 |
| DE | .3 | 95 | 0 | 20 | 3 | 69 | 41 |
| DC | .3 | 96 | 0 | 19 | 7 | 71 | 7 |
| FL | 4.5 | 91 | 1 | 11 | 4 | 0 | 0 |
| GA | 2.4 | 88 | 6 | 15 | 4 | 34 | 0 |
| HI | .4 | 95 | .02 | 22 | 4 | 68 | 43 |
| ID | .4 | 93 | 7 | 19 | 2 | 62 | 48 |
| IL | 4.9 | 95 | 2 | 11 | 7 | 0 | 0 |
| IN | 2.3 | 93 | 5 | 13 | 4 | 33 | 0 |
| IA | 1.3 | 97 | 11 | 14 | 2 | 62 | 38 |
| KS | 1.0 | 95 | 10 | 17 | 4 | 62 | 0 |
| KY | 1.6 | 87 | 7 | 17 | 3 | 41 | 0 |
| LA | 1.8 | 90 | 3 | 18 | 4 | 49 | 5 |
| ME | .5 | 93 | 3 | 13 | 2 | 44 | 0 |
| MD | 2.0 | 96 | 4 | 15 | 4 | 60 | 0 |
| MA | 2.6 | 96 | 0 | 10 | 4 | 5 | 0 |
| MI | 3.9 | 96 | 2 | 11 | 4 | 13 | 0 |
| MN | 2.0 | 97 | 7 | 13 | 2 | 59 | 0 |
| MS | 1.0 | 84 | 8 | 19 | 3 | 42 | 26 |
| MO | 2.0 | 95 | 5 | 13 | 2 | 42 | 0 |
| MT | .4 | 93 | 9 | 18 | 3 | 60 | 31 |
| NE | .7 | 96 | 9 | 19 | 4 | 70 | 42 |
| NV | .5 | 91 | 3 | 21 | 4 | 58 | 22 |
| NH | .5 | 94 | 0 | 14 | 3 | 51 | 0 |
| NJ | 3.2 | 95 | 0 | 11 | 7 | 11 | 0 |
| NM | .5 | 86 | 5 | 20 | 5 | 45 | 0 |
| NY | 7.0 | 92 | 1 | 8 | 6 | 0 | 0 |
| NC | 2.7 | 89 | 2 | 14 | 4 | 23 | 0 |
| ND | .3 | 96 | 14 | 18 | 3 | 70 | 53 |
| OH | 4.5 | 94 | .8 | 11 | 5 | 0 | 0 |
| OK | 1.4 | 92 | 5 | 17 | 5 | 54 | 0 |
| OR | 1.3 | 93 | 4 | 18 | 4 | 59 | 0 |
| PA | 4.8 | 96 | 2 | 10 | 4 | 0 | 0 |
| RI | .4 | 95 | 0 | 17 | 3 | 61 | 29 |
| SC | 1.3 | 87 | 10 | 17 | 4 | 44 | 0 |
| SD | .3 | 94 | 14 | 17 | 3 | 64 | 41 |
| TN | 2.0 | 89 | 7 | 17 | 3 | 47 | 8 |
| TX | 6.8 | 90 | 6 | 11 | 4 | 0 | 0 |
| UT | .7 | 95 | 3 | 17 | 4 | 59 | 9 |
| VT | .2 | 93 | 0 | 16 | 2 | 59 | 20 |
| VA | 2.4 | 93 | 13 | 15 | 4 | 50 | 0 |
| WA | 1.9 | 94 | 4 | 18 | 4 | 61 | 7 |
| WV | .7 | 90 | 6 | 17 | 4 | 46 | 4 |
| WI | 2.1 | 97 | 11 | 14 | 2 | 63 | 0 |
| WY | .3 | 92 | 7 | 24 | 4 | 66 | 44 |

3. Performance of $\bar{x}_{KS}$, $\bar{x}_{US}$, $\bar{x}_{KC}$, and $\bar{x}_{UC}$

Relative to the Optimum Area/List Frame Estimator

| | MSE Reduction | | Telephone Allocation | |
|---|---|---|---|---|
| | TBIAS = 0% | TBIAS = 5% | TBIAS = 0% | TBIAS = 5% |
| $\bar{x}_{KS}$ | 7% | 2% | 24% | 3% |
| $\bar{x}_{US}$ | 6% | 1% | 23% | 3% |
| $\bar{x}_{KC}$ | 3% | 1% | 31% | 1% |
| $\bar{x}_{UC}$ | 2% | 0% | 29% | 1% |

## REFERENCES


Bosecker, Raymond R. and Ford, Barry L. (1976). Multiple Frame Estimation with Stratified Overlap Domain," Statistical Reporting Service, U.S. Department of Agriculture.

Casady, Robert J., and Sirken, Monroe G. (1980). "A Multiplicity Estimator for Multiple Frame Sampling," Proceedings of the American Statistical Association, Social Statistics Section.

__________, Snowden, Cecilia B., and Sirken, Monroe G. (1981). "A Study of Dual Frame Estimators for the National Health Interview Study," Proceedings of the American Statistical Association, Social Statistics Section.

Cochran W.G. (1977). Sampling Techniques, John Wiley and Sons, New York.

Raj, D. (1966). "Some Remarks on a Simple Procedure of Sampling Without Replacement," Journal of the American Statistical Association, 61.

Hansen, Morris H., Hurwitz, William N. and Madow, William G. (1953). Sample Survey Methods and Theory, John Wiley and Sons, New York.

Hartley, H.O. (1962). "Multiple Frame Sample Surveys" Proceedings of the American Statistical Association, Social Statistics Section.

Hartley, H.O. (1974). "Multiple Frame Methodology and Selected Applications," Sankya.

Lund, Richard E. (1968). "Estimators in Multiple Frame Surveys," Proceedings of the American Statistical Association, Social Science Section.

McGowan, Howard R. (1982). "Telephone Ownership in the National Crime Survey," Unpublished manuscript.

Thornberry, Owen T., Jr. and Massey, James T. (1978). "Correcting for Undercoverage Bias in Random Digit Dialed National Health Surveys," _Proceedings of the American Statistical Association_, _Section on Survey Research Methods_.

Waksberg, Joseph (1978). "Sampling Methods for Random Digit Dialing," _Journal of the American Statistical Association_, 73, 40-46.