Journal of Official Statistics
Vol. 6. No. 4, 1990, pp. 437–449

Prediction Theory Approach to Multistage Sampling When Cluster Sizes Are Unknown

Elizabeth J. Kelly[1] *and William G. Cumberland*[2]

Abstract: A model for two-stage cluster sampling when sample cluster sizes are unknown is used to derive an optimal estimator for the population total and to determine robust sampling strategies. In an empirical study using a real population, comparisons were made between the model-based estimator and conventional estimators. The results favored the new model-based estimator over traditional estimators derived from randomization theory. In the empirical study robust sampling strategies suggested by the theory reduced biases, improved efficiency, and decreased the frequencies of large errors.

Key words: Model-based estimation; robust estimators; two-stage cluster sampling; prediction; empirical study; bias.

1. Introduction

Model-based inference for finite populations has provided valuable insight into the behavior of conventional estimators, and led to the introduction of new, robust variance estimators (Royall 1976, 1986; Royall and Cumberland 1978, 1981a, 1981b; Cumberland and Royall 1981). In this paper, we use prediction theory to study two-stage cluster sampling when cluster sizes are unknown. A superpopulation model discussed by Royall (1986) is used to develop criteria for evaluating sampling strategies, i.e., sampling designs and estimators. An empirical study, using 1970 and 1980 census data for Los Angeles and surrounding counties, corroborates the prediction theory results.

[1] Los Alamos National Laboratory, Los Alamos, New Mexico, U.S.A. [2] Department of Biostatistics, School of Public Health, University of California, Los Angeles, California, U.S.A.

2. Two-Stage Sampling with Unknown Cluster Sizes

The population is divided into primary units (or clusters) each containing M_i secondary units. The outcome variable y is measured on the secondary units. The problem is one of estimating the population total, $T = \Sigma_{i=1}^{N}\Sigma_{j=1}^{M_i} y_{ij}$, from the sampled secondary units.

In the two-stage sampling framework a sample, s, of n clusters is taken and a subsample, s_i, of m_i secondary units is drawn from each sampled cluster. We assume all sampling is without replacement. The situation considered here is commonly encountered, where the M_i are known only for the sampled clusters, but there exist related auxiliary variables, X_i, that are known for all clusters. The total number of sampled secondary units is $m_s = \Sigma_{i\in s} m_i$, while the sample means of the y_{ij} for each cluster are given by $\bar{y}_{s_i} = \Sigma_{j\in s_i} y_{ij}/m_i$, $i = 1, 2, \ldots, n$.

Letting r denote the clusters that are not in the first-stage sample, the following examples illustrate the notational convention used for summations of any variable: $X_r = \Sigma_{i \in r} X_i$,

$$\bar{X}_s = n^{-1} \sum_{i \in s} X_i, \quad (\overline{X^2})_s = n^{-1} \sum_{i \in s} X_i^2,$$

$$Y_i = \sum_{j=1}^{M_i} y_{ij}, \qquad M = \sum_{i=1}^{N} M_i.$$

The quantity $f = n/N$ is the first-stage sampling fraction and $f_i = m_i/M_i$ is the second-stage sampling fraction.

3. Conventional Estimators for the Population Total

A sampling design used by many large surveys consists of selecting the clusters with probability proportional to a size measure (X_i), and the secondary units with simple random sampling (PPS–SRS). In this case, the Horvitz–Thompson estimator is generally used to estimate the population total, $\hat{T}_P = Nn^{-1} \Sigma_{i \in s} (M_i \bar{y}_{s_i} \bar{X}/X_i)$. Another common sampling strategy consists of selecting both the clusters and the secondary units by simple random sampling (SRS–SRS) and then employing the ratio estimator, $\hat{T}_R = Nn^{-1} \Sigma_{i \in s} (M_i \bar{y}_{s_i}) \bar{X}/\bar{X}_s$. Traditionally $\hat{T}_P$ and $\hat{T}_R$ have been the estimators chosen for populations where the Y_i are thought to be approximately proportional to the X_i. When no such X_i exist, then for SRS–SRS designs the "unbiased" estimator, $\hat{T}_U$, has been suggested (Cochran 1977): $\hat{T}_U = Nn^{-1} \Sigma_{i \in s} M_i \bar{y}_{s_i}$. The Horvitz–Thompson estimator, $\hat{T}_P$, is unbiased with respect to a PPS–SRS plan and $\hat{T}_U$ is unbiased with respect to an SRS–SRS plan. The ratio estimator, $\hat{T}_R$, is biased with respect to SRS–SRS; however, the bias has been shown to be negligible for large n (Cochran 1977). Although these estimators were developed under traditional sampling theory, insight into their behavior can be gained when they are evaluated as estimators for the population total using prediction theory.

4. Prediction Theory Estimators

The prediction approach to finite population sampling treats the y_{ij} as realizations of random variables Y_{ij}. In this application, cluster sizes are not known for nonsampled clusters; therefore, the M_i are also treated as realizations of random variables. The superpopulation model proposed by Royall (1986) describes a population where cluster size is proportional to a previous size measure and cluster totals are increasing linearly conditionally with cluster size. The Y_{ij} are assumed to be correlated within clusters but independent between clusters. Denoting conditional expectations, variances, and covariances by E^*, Var*, and Cov*, Royall's model was the following:

MODEL M_μ:

i. $E(M_i) = \beta X_i \quad i = 1, 2, \ldots, N$

ii. $\mathrm{Var}(M_i) = \tau^2 X_i$
and $\mathrm{Cov}(M_i, M_j) = 0 \quad i \neq j$

iii. $\Pr(M_i < 2) = 0$

iv. $E^*(Y_{ij}) = \mu \quad j = 1, 2, \ldots, M_i$

v. $\mathrm{Var}^*(Y_{ij}) = \sigma_i^2$

vi. $\mathrm{Cov}^*(Y_{ij}, Y_{kl}) = \begin{cases} \rho_i \sigma_i^2 & i = k, \quad j \neq l \\ 0 & i \neq k \end{cases}$

vii. $n > 2$.

Royall (1986) used this model to derive robust estimators for the error variances of the ratio and Horvitz–Thompson estimators. We used the model to derive an optimal model-based estimator for the population total and an estimator of its variance.

The parameters β, μ, and τ^2 are constants. We consider only designs where $m_i \geq 2$

and, if $m_i > M_i$, we take $m_i = M_i$. We assume ρ_i is nonnegative; this is not a strong restriction since it can be shown that $\rho_i > -1/(M_i - 1)$. In many of the analyses that follow, the restrictions $\rho_i = \rho$ and $\sigma_i^2 = \sigma^2$ are made; the model with these restrictions is denoted M_μ'.

After the sample is observed, the population total can be written as the sum of the observed and unobserved variables

$$T = \sum_{i\in s}\sum_{j\in s_i} y_{ij} + \sum_{i\in s}\sum_{j\in r_i} y_{ij} + \sum_{i\in r}\sum_{j=1}^{M_i} y_{ij}$$

where r_i is the set of subunits not included in the sample s_i. The first term is known and the problem of estimating T is equivalent to that of predicting the total for the unobserved y_{ij}

$$\sum_{i\in s}\sum_{j\in r_i} y_{ij} + \sum_{i\in r}\sum_{j=1}^{M_i} y_{ij}.$$

The best linear unbiased estimator for T, $\dot{T}_{\text{BLU}}$, is found by adding the observed total to the BLU predictor of the unobserved total. This technique was used by Royall (1976) for the case where the M_i were known. The BLU predictor can be found by exploiting a result from prediction theory, which states that given a $k + p$ random vector $\mathbf{Z}$ with mean $\mathbf{U}$ and covariance Σ

$$\mathbf{Z} = \begin{bmatrix}\mathbf{Z}_p\\ \mathbf{Z}_k\end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix}\mathbf{U}_p\\ \mathbf{U}_k\end{bmatrix},$$

$$\Sigma = \begin{bmatrix}\Sigma_{pp} & \Sigma_{pk}\\ \Sigma_{kp} & \Sigma_{kk}\end{bmatrix}$$

the best linear unbiased predictor of $\mathbf{Z}_p$ given $\mathbf{Z}_k$ is

$$\dot{\mathbf{Z}}_p = \mathbf{U}_p + \Sigma_{pk}\Sigma_{kk}^{-1}(\mathbf{Z}_k - \mathbf{U}_k). \qquad (4.1)$$

We use this theorem by properly defining $\mathbf{Z}$. After conditioning on the observed sample s, r, m_i, and s_i are fixed. We define $\mathbf{Z}$ as the $N + 2n$ random vector

$$\mathbf{Z} = \begin{bmatrix}\mathbf{Z}_p\\ \mathbf{Z}_k\end{bmatrix}, \quad \mathbf{Z}_p = \begin{bmatrix}\mathbf{Z}_{\text{I}}\\ \mathbf{Z}_{\text{II}}\end{bmatrix}$$

and

$$\mathbf{Z}_k = \begin{bmatrix}\mathbf{Z}_{\text{III}}\\ \mathbf{Z}_{\text{IV}}\end{bmatrix}$$

where

$$\mathbf{Z}_{\text{I}} = \left(\left(\sum_{j\in r_i} Y_{ij}\right)\right) \quad i \in s$$

$$\mathbf{Z}_{\text{II}} = \left(\left(\sum_{j=1}^{M_i} Y_{ij}\right)\right) \quad i \in r$$

$$\mathbf{Z}_{\text{III}} = \left(\left(\sum_{j\in s_i} Y_{ij}\right)\right) \quad i \in s$$

$$\mathbf{Z}_{\text{IV}} = ((M_i)) \quad i \in s.$$

Then, $\mathbf{Z}$ has covariance matrix

$$\Sigma = \begin{bmatrix}\text{V}_{\text{I}} & 0 & \text{V}_{\text{I,III}} & \text{V}_{\text{I,IV}}\\ 0 & \text{V}_{\text{II}} & 0 & 0\\ \text{V}_{\text{I,III}} & 0 & \text{V}_{\text{III}} & 0\\ \text{V}_{\text{I,IV}} & 0 & 0 & \text{V}_{\text{IV}}\end{bmatrix}$$

where V_{I} is the $n \times n$ covariance matrix of $\mathbf{Z}_{\text{I}}$, $\text{V}_{\text{I,III}}$ is the $n \times n$ covariance matrix of $\mathbf{Z}_{\text{I}}$ and $\mathbf{Z}_{\text{III}}$, V_{II} is the $N - n \times N - n$ covariance matrix of $\mathbf{Z}_{\text{II}}$, etc. The expectations are given by

$$E(\mathbf{Z}_{\text{I}}) = ((\mu\beta X_i - \mu m_i)) \quad i \in s$$

$$E(\mathbf{Z}_{\text{II}}) = ((\mu\beta X_i)) \quad i \in r$$

$$E(\mathbf{Z}_{\text{III}}) = ((\mu m_i)) \quad i \in s$$

$$E(\mathbf{Z}_{\text{IV}}) = ((\beta X_i)) \quad i \in s.$$

The matrix $\text{V}_{\text{I,III}}$ is diagonal with elements

$$\text{Cov}\left(\sum_{j\in r_i} Y_{ij}, \sum_{j\in s_i} Y_{ij}\right) = (\beta X_i - m_i)m_i\rho_i\sigma_i^2.$$

The matrix $\text{V}_{\text{I,IV}}$ is diagonal with elements

$$\text{Cov}\left(\sum_{j\in r_i} Y_{ij}, M_i\right) = \mu\tau^2 X_i.$$

The matrix V_{III} is diagonal with elements

$$\text{Var}\left(\sum_{i\in s_i} Y_{ij}\right) = m_i[(1-\rho_i)\sigma_i^2 + m_i\rho_i\sigma_i^2]$$

and V_{IV} is diagonal with elements $\tau^2 X_i$. Applying theorem (4.1) gives the predictors

$$\dot{\mathbf{Z}}_i = (M_i - m_i)\mu + (\beta X_i - m_i)\frac{m_i\rho_i\sigma_i^2}{(1-\rho_i)\sigma_i^2 + m_i\rho_i\sigma_i^2} \times (\bar{y}_{s_i} - \mu) \quad i \in s$$

and

$$\dot{\mathbf{Z}}_i = \mu\beta X_i \quad i \in r.$$

Let $w_i = m_i\rho_i/[(1-\rho_i) + m_i\rho_i]$, then the BLU estimator for T is

$$\dot{T}_{\text{BLU}} = \sum_{i\in s}\sum_{j\in s_i} y_{ij} + \sum_{i\in s}(M_i - m_i)\mu + \sum_{i\in s}(\beta X_i - m_i)w_i(\bar{y}_{s_i} - \mu) + \beta\mu X_r.$$

Substituting M_i for βX_i in the third term, gives a nonlinear estimator, $\dot{T}_{\text{NL}}$

$$\dot{T}_{\text{NL}} = \sum_{i\in s}\sum_{j\in s_i} y_{ij} + \sum_{i\in s}(M_i - m_i) \times [w_i\bar{y}_{s_i} + (1-w_i)\mu] + \beta\mu X_r.$$

When the parameters β and μ are unknown, they can be estimated by their BLU estimators

$$\hat{\beta} = \bar{M}_s/\bar{X}_s \quad \text{and} \quad \hat{\mu} = \sum_{i\in s} u_i\bar{y}_{s_i}$$

where

$$u_i = \frac{m_i/[(1-\rho_i)\sigma_i^2 + m_i\rho_i\sigma_i^2]}{\Sigma_{i\in s}\, m_i/[(1-\rho_i)\sigma_i^2 + m_i\rho_i\sigma_i^2]}.$$

The BLU and nonlinear estimators with $\hat{\beta}$ and $\hat{\mu}$ substituted for β and μ are denoted $\ddot{T}_{\text{BLU}}$ and $\ddot{T}_{\text{NL}}$. With respect to M_μ, $\ddot{T}_{\text{NL}}$ is also unbiased and comparing MSEs we find

$$E(\ddot{T}_{\text{BLU}} - T)^2 - E(\ddot{T}_{\text{NL}} - T)^2 = \sum_{i\in s}[\text{Var}(M_i) - \text{Var}(\hat{\beta}X_i)] \times w_i^2[\text{Var}(\bar{y}_{s_i}) - \text{Var}(\hat{\mu})].$$

The sum is nonnegative, therefore, $\ddot{T}_{\text{NL}}$ has smaller MSE than $\ddot{T}_{\text{BLU}}$. Deriving this result takes some effort. The details are supplied in Cohen (1984). We concentrate on $\ddot{T}_{\text{NL}}$ since it has smaller MSE than $\ddot{T}_{\text{BLU}}$.

The estimator $\ddot{T}_{\text{NL}}$ depends on u_i and w_i, which depend on $\rho_i\sigma_i^2$ and $(1-\rho_i)\sigma_i^2$, and are generally unknown. For the case $\rho_i = \rho$ and $\sigma_i^2 = \sigma^2$, Rustagi (1978) notes that M_μ' is equivalent to a one-way random effects model: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m_i$; where $\rho\sigma^2 = \text{Var}(\alpha_i) = \sigma_\alpha^2$ and $(1-\rho)\sigma^2 = \text{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$. Unbiased estimators for σ_ε^2 and σ_α^2 are

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i\in s}\sum_{j\in s_i}\frac{(y_{ij} - \bar{y}_{s_i})^2}{(m_i - 1)}$$

and

$$\hat{\sigma}_\alpha^2 = S_b^2 - \frac{1}{n}\sum_{i\in s}\frac{1}{m_i}\sum_{j\in s_i}\frac{(y_{ij} - \bar{y}_{s_i})^2}{(m_i - 1)}$$

where

$$S_b^2 = \frac{1}{(n-1)}\sum_{i\in s}(\bar{y}_{s_i} - \dot{y})^2,$$

$$\dot{y} = \frac{1}{n}\sum_{i\in s}\bar{y}_{s_i}.$$

These are the unweighted sum of squares estimators (USS) from random effects analysis. In the following discussion, the USS estimators are substituted for $(1-\rho)\sigma^2$ and $\rho\sigma^2$ in $\ddot{T}_{\text{NL}}$ and the resulting estimator for the population total is denoted $\hat{T}_{\text{NL}}$.

5. Model-Based Analysis

5.1. *Biases under model M_μ*

When expectation is taken with respect to M_μ, $\hat{T}_P$ and $\hat{T}_R$ are unbiased and the model-

based estimator $\hat{T}_{NL}$ is asymptotically unbiased (m_i and n large). The "unbiased" estimator $\hat{T}_U$ is biased

$$E(\hat{T}_U - T) = N\beta\mu(\bar{X}_s - \bar{X}).$$

5.2. *Model failure – misspecified expectation*

The superpopulation model is a working model that describes the gross structure of many real populations. Deviations from this or any other model are to be expected. An important part of the prediction approach to estimation is to evaluate estimators under model failure and determine strategies that are robust to such deviations.

In the case of single-stage sampling (Royall and Cumberland 1981a, 1981b; Cumberland and Royall 1981) and two-stage sampling when cluster size is known (Royall 1976), certain model failures caused conditional biases. Valliant (1987) showed that the use of stratification and reasonably large samples was not necessarily sufficient to remove these conditional biases. It is not surprising then to find that conditional biases exist for certain failures of the assumptions regarding M_μ. For example, when cluster size is not proportional to the previous size measure, but can be described by a polynomial with an intercept and quadratic term

$$E(M_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$$

then, under the remaining assumptions of M_μ' the conditional biases are

$$E(\hat{T}_P - T) = N\mu\beta_0[(\overline{X^{-1}})_s\bar{X} - 1] + N\mu\beta_2[\bar{X}\bar{X}_s - (\overline{X^2})]$$

$$E(\hat{T}_R - T) = N\mu\beta_0\left[\frac{\bar{X} - \bar{X}_s}{\bar{X}_s}\right] + N\mu\beta_2\left[\frac{\bar{X}(\overline{X^2})_s - \bar{X}_s(\overline{X^2})}{\bar{X}_s}\right]$$

$$E(\hat{T}_U - T) = N\mu\beta_1[\bar{X}_s - \bar{X}] + N\mu\beta_2[(\overline{X^2})_s - (\overline{X^2})]$$

and

$$E(\hat{T}_{NL} - T) \approx N\mu\beta_0\left[\frac{\bar{X} - \bar{X}_s}{\bar{X}_s}\right] + N\mu\beta_2\left[\frac{\bar{X}(\overline{X^2})_s - \bar{X}_s(\overline{X^2})}{\bar{X}_s}\right].$$

(Note that $E(\hat{T}_{NL} - T)$ is an asymptotic result except when ρ and σ^2 are known.) These biases depend on sample characteristics and the theory indicates that the estimators can be protected from such biases if restrictions are placed on the sampled X_i. The restrictions for $\hat{T}_P$ are $(\overline{X^{-1}})_s = 1/\bar{X}$ and $\bar{X}\bar{X}_s = (\overline{X^2})$ (π-balance (Cumberland and Royall 1981)). The restrictions for $\hat{T}_R$, $\hat{T}_U$, and $\hat{T}_{NL}$ are $\bar{X} = \bar{X}_s$ and $(\overline{X^2})_s = (\overline{X^2})$ (balance on the first and second moments (Royall and Herson 1973)). These results support previous findings from single-stage sampling (Royall and Cumberland 1981a and Cumberland and Royall 1981), and indicate that sampling techniques that force balance can protect against certain biases in multistage sampling when cluster size is unknown.

5.3. *Variance estimators*

Royall (1986) introduced robust variance estimators based on the sum of squares of the residuals, which are unbiased and consistent under the restricted model M_μ' and, under mild conditions on how the sample and population grow as n and $N \to \infty$ and $f \to 0$, these variance estimators remain consistent under the general model M_μ. Royall (1986) noted that a general expression for the estimator of the population total is $\hat{T} = N\,\Sigma_{i\in s}\, u_i \hat{Y}_i/n_i$ where $u_i = \bar{X}/\bar{X}_s$, $\bar{X}/X_i$, and 1 given $\hat{T}_R$, $\hat{T}_P$, and $\hat{T}_U$, respectively. Royall (1986) derived a general

expression for the model-based robust variance estimators

$$v_0 = \left(\frac{N}{n}\right)^2 \sum_{i\in s} u_i(u_i - n/N)\hat{v}_i + \frac{N}{n}\sum_{i\in s} u_i M_i(1 - f_i)S_i^2/f_i + \left(\sum_{i=1}^{N} X_i^2 - \frac{N}{n}\sum_{i\in s} u_i X_i^2\right)\theta_2$$

where $\hat{v}_i$ is an unbiased estimate of $\mathrm{Var}(\hat{Y}_i)$ based on the sum of squares of the residuals, $\hat{Y}_i - (\hat{T}/N\bar{X})X_i$, θ_2 is an unbiased estimate of $\beta^2\rho\sigma^2$, and $S_i^2 = \Sigma_{i\in s_i}(y_{ij} - \bar{y}_{s_i})^2/(m-1)$. Substituting the appropriate values of u_i gives the model-based variance estimators v_R, v_U, and v_P for $\hat{T}_R$, $\hat{T}_U$, and $\hat{T}_P$. The variance estimator for $\hat{T}_{\mathrm{NL}}$ is derived from the MSE of $\ddot{T}_{\mathrm{NL}}$ under the model M'_μ. If $m_i = m$, then

$$\begin{aligned} E(\ddot{T}_{\mathrm{NL}} - T)^2 &= \sum_{i\in s}(M_i - m)^2 \\ &\quad \times [w_i^2\,\mathrm{Var}(\bar{y}_s) + (1 - w_i^2)\,\mathrm{Var}(\hat{\mu})] \\ &\quad + (\mathrm{Var}(\hat{\beta}) + \beta^2)\,\mathrm{Var}(\hat{\mu})\,X_r^2 \\ &\quad + (M_s - m_s)\sigma_\varepsilon^2 + \sum_{i\in s}(M_i - m_s)\sigma_\alpha^2 \\ &\quad + \beta X_r\sigma_\varepsilon^2 + [\tau^2 X_r + \beta^2(X^2)_r]\sigma_\alpha^2 \\ &\quad - 2\left[\sum_{i\in s}(M_i - m)^2 w_i\sigma_\alpha^2 \right. \\ &\quad \left. + \sum_{i\in s}(M_i - m)\,(\bar{M}_s - m)(1 - w_i)\sigma_\alpha^2\right] \\ &\quad + 2(M_s - m)\beta X_r\sigma_\varepsilon^2/m \\ &\quad + \mu^2 X_r X\,\mathrm{Var}(\hat{\beta}). \end{aligned}$$

The variance estimator, v_{NL}, is found by substituting the following unbiased estimators for the unknown parameters

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i\in s}\sum_{j\in s_i}\frac{(y_{ij} - \bar{y}_{s_i})^2}{(m-1)}$$

$$\hat{\sigma}_\alpha^2 = S_b^2 - \frac{1}{nm}\hat{\sigma}_\varepsilon^2$$

$$S_b^2 = \frac{1}{n-1}\sum_{i\in s}(\bar{y}_{s_i} - \dot{y})^2, \quad \dot{y} = \frac{1}{n}\sum_{i\in s}\bar{y}_{s_i}$$

$$\hat{w}_i = m\hat{\sigma}_\alpha^2/(m\hat{\sigma}_\alpha^2 + \hat{\sigma}_\varepsilon^2)$$

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \sum_s (M_i - \hat{\beta}X_i)^2 \Big/ \left(X_s^2 - \sum_{i\in s} X_i^2\right)$$

$$\widehat{\mathrm{Var}}(\bar{y}_{s_i}) = \frac{1}{n-1}\sum_{i\in s}(\bar{y}_{s_i} - \hat{\mu})^2$$

$$\widehat{\mathrm{Var}}(\hat{\mu}) = 1/n\ \widehat{\mathrm{Var}}(\bar{y}_{s_i})$$

$$\widehat{\tau^2} = X_s\,\widehat{\mathrm{Var}}(\hat{\beta})$$

$$\widehat{\beta^2} = \hat{\beta}^2 - \widehat{\mathrm{Var}}(\hat{\beta}) \quad \text{and}$$

$$\widehat{\mu^2} = \hat{\mu}^2 - \widehat{\mathrm{Var}}(\hat{\mu}).$$

6. Empirical Study

The goal of the empirical study was to test the theoretical results against a real data set that could not be described exactly by a model. Such an investigation provides further insight into the problems of multistage sampling and estimation, and tests the robustness of the model-based theoretical results in situations where the model failure is more complex than that described in Section 5.2.

In this study, we repeatedly drew two-stage samples from block statistics for outlying areas of Los Angeles County, and all of Ventura and Orange Counties using the 1970 and 1980 census data. Primary units were census tracts with previous size measures, X_i, taken from the 1970 counts of the numbers of blocks in each tract, and current size measures, M_i, taken from the 1980 counts. The outcome measure, y_{ij}, was the 1980 block population. If a census tract had fewer than twenty blocks in the 1970 census, then it was combined with its nearest neighbor. This process continued until the resulting tract had twenty or more blocks. After this adjustment, there were 420 census tracts. There were 23,001 blocks in 1970;

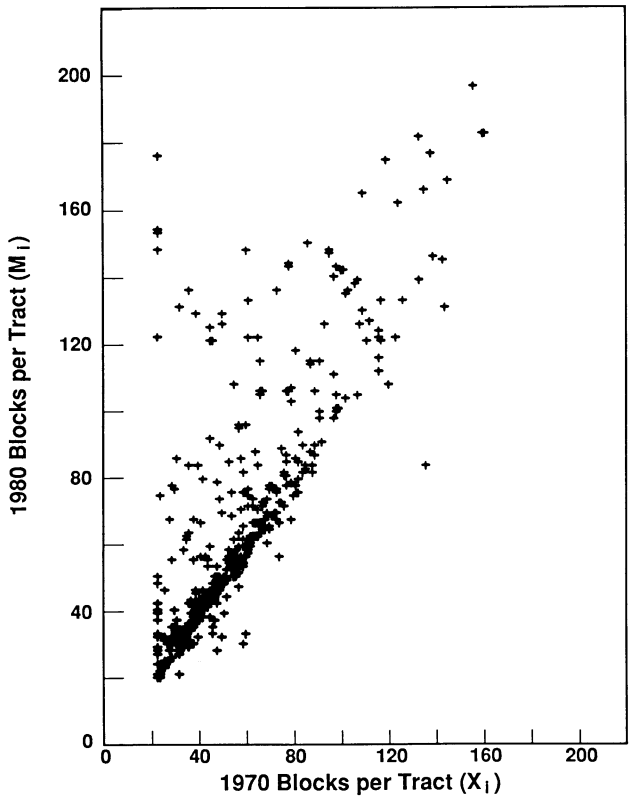

Fig. 1. The current size measures, 1980 Blocks per Tract (M_i), are plotted against the previous size measures, 1970 Blocks per Tract (X_i). This plot supports the assumption that the M_i are increasing linearly with the X_i.

29,102 in 1980, and the total population in 1980 was 4,045,074. Figure 1 is a plot of the current size measure, 1980 blocks per tract (M_i), versus the previous size measure, 1970 blocks per tracts (X_i), and supports the assumption that M_i was increasing linearly with X_i. Finding a data set that would demonstrate this condition was the motivation for choosing the rapidly growing Los Angeles suburbs, and Orange and Ventura Counties. Figure 2 is a plot of the 1980 population per tract, Y_i, versus M_i and supports the assumption that Y_i increased linearly with M_i. These plots indicate that the model provides a reasonable description of this population. However, the model is not perfect. Although a test for curvature was not significant, Figure 2 indicates that there may be some upward curvature in the data. Figure 1 appears to have an intercept term and many of the small clusters show tremendous growth. It is these departures from the straight line through the origin assumptions that necessitate robust sampling strategies (5.2).

6.1. *Sampling strategies*

We took 450 replications for each of five two-stage sampling plans. Forty-two first-stage clusters were drawn using simple random sampling (SRS), approximately balanced SRS (b-SRS), basket sampling (Wallenius 1973), probability proportional

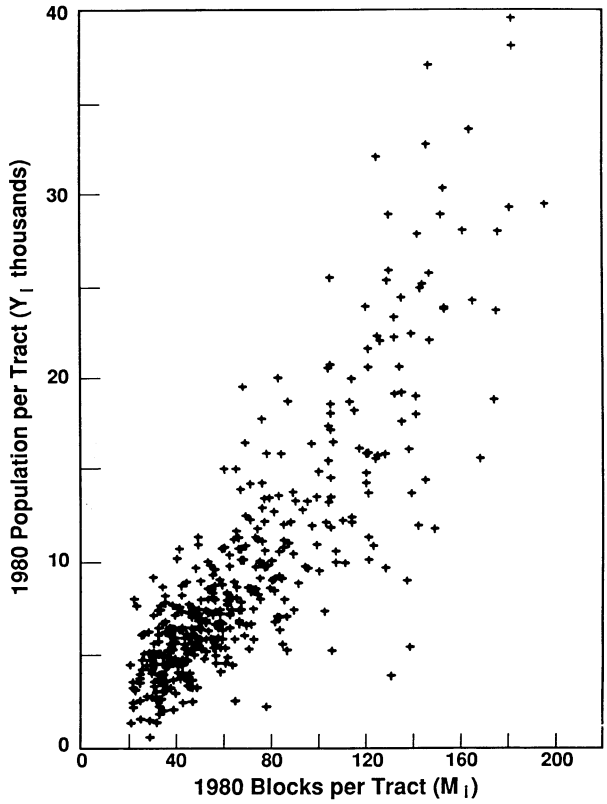

Fig. 2. The 1980 Population per Tract (Y_i) is plotted against the 1980 Blocks per Tract (M_i). This plot supports the assumption that the Y_i are increasing linearly with the M_i.

to size (PPS) sampling, and approximately PPS balanced sampling (b-PPS). For each first-stage sample, twenty second-stage units were selected by simple random sampling. The balance definitions were those given in Section 5.2. The approximate balance requirements were such that 5% of the samples were accepted. The PPS plan consisted of a random permutation of the first-stage units followed by a random start systematic sample with probabilities proportional to size (X_i). Basket sampling (Wallenius 1973) consisted of forming 10 groups of 42 primary units having minimum differences between the sums of the group size measures (X_i) and selecting one of the groups randomly. The resulting samples were extremely well balanced having sample moments close to population moments.

6.2. *Results*

Under PPS sampling (as would be expected when the data deviate from the model assumptions) $\hat{T}_R$, $\hat{T}_U$, and $\hat{T}_{\mathrm{NL}}$ performed poorly, demonstrating large biases and MSEs. The estimator $\hat{T}_P$ was equally bad with SRS or basket sampling, therefore, these sampling strategies were not considered. The estimator $\hat{T}_U$, as the theory predicted and practice has shown, performed poorly, having extremely large conditional biases under SRS sampling. Since $\hat{T}_U$ was equivalent to $\hat{T}_R$ under balanced sampling plans, this estimator was not considered further.

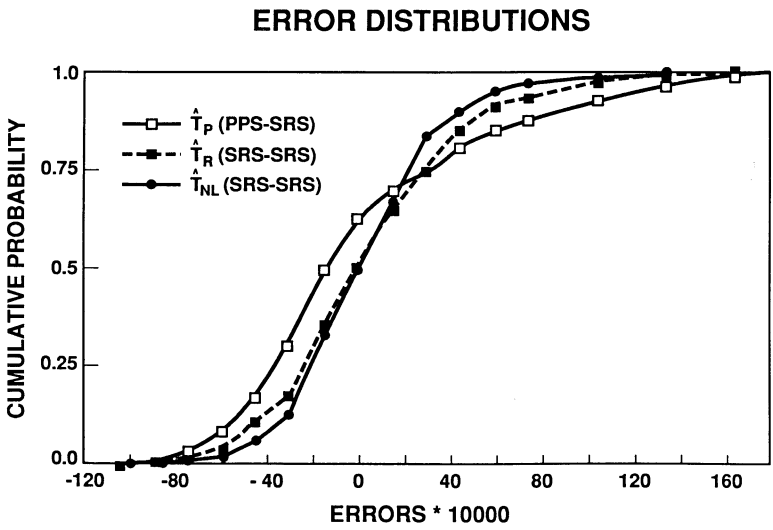

Fig. 3. The cumulative error distributions for $\hat{T}_P$ (PPS–SRS), $\hat{T}_R$ (SRS–SRS), and $\hat{T}_{NL}$ (SRS–SRS) show that $\hat{T}_{NL}$ (SRS–SRS) performed best on the real data – less variability and less likely to have large errors.

Figure 3 shows the cumulative distributions of the errors for $\hat{T}_P$ under PPS–SRS, and $\hat{T}_R$ and $\hat{T}_{NL}$ under SRS–SRS. The estimator $\hat{T}_{NL}$ had smaller errors and less variability than the conventional estimators. For example, the frequency of errors less than $-500{,}000$ and greater than 600,000 was 30% for $\hat{T}_P$, 18% for $\hat{T}_R$, and 10% for $\hat{T}_{NL}$. The relative efficiencies of $\hat{T}_P$ and $\hat{T}_R$ to $\hat{T}_{NL}$ (ratio of MSEs) were 0.56 and 0.74, respectively. Note that $\hat{T}_R$ with SRS–SRS sampling performed better than $\hat{T}_P$ with PPS–SRS sampling. The relative efficiency of $\hat{T}_P$ to $\hat{T}_R$ was 0.69. The error distributions of all of the estimators were positively skewed.

In Figures 4, 5, and 6; bias, $(\text{MSE})^{1/2}$, and model-based variance estimators $(v_P)^{1/2}$, $(v_R)^{1/2}$, $(v_{NL})^{1/2}$ (Section 5.3) were plotted as functions of sample characteristics. For $\hat{T}_{NL}$ and $\hat{T}_R$, the 450 replications were ordered according to increasing $\bar{X}_s$ and divided into 10 groups of 45 samples. For each group bias, $(\text{MSE})^{1/2}$, and $(v)^{1/2}$ were calculated and plotted versus the group average of the $\bar{X}_s$. For $\hat{T}_P$ the same procedure was followed except $(\overline{X^{-1}})_s$ was used instead of $\bar{X}_s$. The plots demonstrated the conditional biases of these estimators predicted by the theory when the model was misspecified (Section 5.2). Similar biases were seen for the ratio and Horvitz–Thompson estimators in single stage sampling (Royall and Cumberland 1981a; Cumberland and Royall 1981) and the combined ratio estimator with stratified SRS (Valliant 1987). The estimator $\hat{T}_P$ had the most severe biases, ranging from $-200{,}000$ to 300,000.

The theory predicted that balanced sampling would reduce or eliminate these biases for certain types of model failure. To investigate this result, $\hat{T}_R$ and $\hat{T}_{NL}$ with b-SRS and basket sampling, and $\hat{T}_P$ with b-PPS sampling were studied. The results showed that approximately balanced samples did indeed reduce conditional biases. For $\hat{T}_R$ and $\hat{T}_{NL}$, basket sampling gave the greatest improvement. Figure 7 shows the cumulative distribution of the errors for $\hat{T}_R$ and $\hat{T}_{NL}$ with basket sampling and $\hat{T}_P$ with b-PPS sampling.

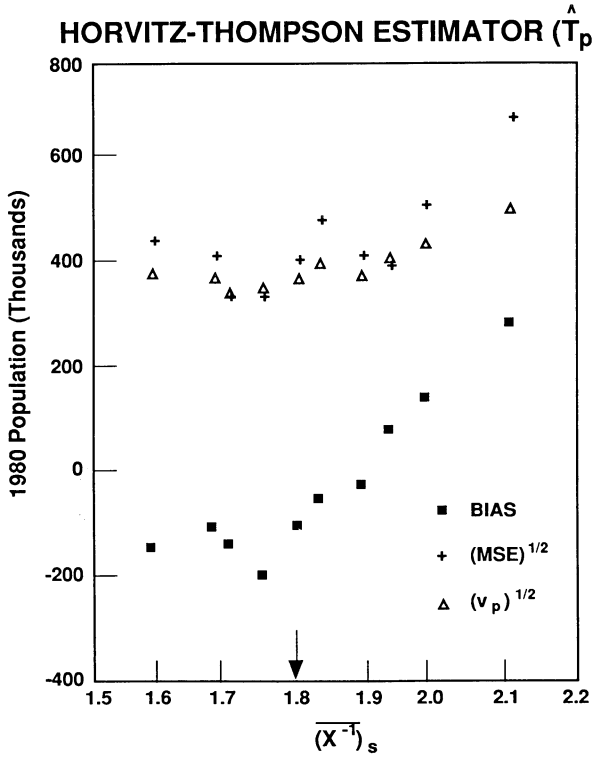

Fig. 4. Bias, $(MSE)^{1/2}$, and $(v_P)^{1/2}$ are plotted as functions of $(\overline{X^{-1}})_s$. The bias plot shows the linear dependency of conditional biases on $(\overline{X^{-1}})_s$, which was predicted by the theory under certain types of model failure. The $(MSE)^{1/2}$ and $(v_P)^{1/2}$ plots show that the robust variance estimator developed by Royall tracks the actual $(MSE)^{1/2}$ fairly well. The arrow marks the π-balance point.

Balanced sampling reduced variability and the probability of large errors as well as the conditional biases. Again, $\hat{T}_{NL}$ looked good, having no errors greater than 750,000 or less than −450,000, while $\hat{T}_R$ had 11% and $\hat{T}_P$ had 22% outside these bounds. The relative efficiencies of $\hat{T}_R$ and $\hat{T}_P$ to $\hat{T}_{NL}$ were 0.51 and 0.36, respectively. Again $\hat{T}_R$ was superior to $\hat{T}_P$. The relative efficiency of $\hat{T}_P$ to $\hat{T}_R$ was 0.7. The error distributions of all three estimators again skewed positively.

The variance estimators were plotted as functions of sample characteristics $\bar{X}_s$, $(\overline{X^{-1}})_s$ in Figures 4, 5, and 6. Conventional variance estimators for $\hat{T}_R$ and $\hat{T}_P$ (Cochran 1977) did not differ significantly from the model-based estimators for this population. Therefore, only the results of the model-based estimators are presented. The estimators, $(v_P)^{1/2}$, $(v_R)^{1/2}$, and $(v_{NL})^{1/2}$, tracked the $(MSE)^{1/2}$ quite well, however, the variance estimator for $\hat{T}_P$ tended to underestimate the variance unless the samples were approximately PPS balanced.

7. Conclusions

The superpopulation model, M_μ, led to a nonlinear estimator for the population total. This estimator ($\hat{T}_{NL}$) had theoretical advantages and, in an empirical study, per-

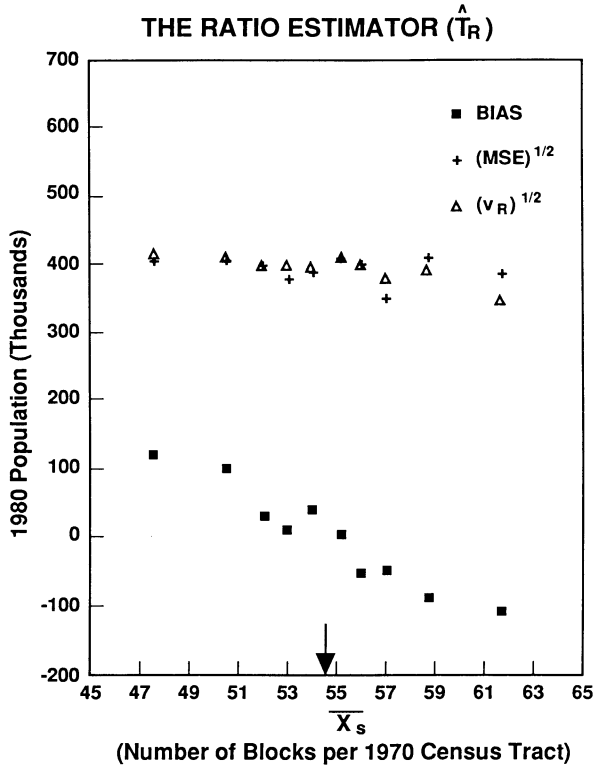

Fig. 5. Bias, $(MSE)^{1/2}$, and $(v_R)^{1/2}$ are plotted as functions of the sample means. The bias plot shows the linear dependency of conditional biases on $\bar{X}_s$, which was predicted by the theory under certain types of model failure. The $(MSE)^{1/2}$ and $(v_R)^{1/2}$ plots show that the robust variance estimator developed by Royall tracks the actual $(MSE)^{1/2}$ quite closely. The arrow marks the balance point.

formed better than the conventional estimators (a Horvitz–Thompson estimator, $(\hat{T}_P)$; and a ratio estimator, $(\hat{T}_R)$). The estimator $\hat{T}_{NL}$ with SRS–SRS sampling had smaller errors and was more efficient than $\hat{T}_R$ with SRS–SRS or $\hat{T}_P$ with PPS–SRS. The estimator $\hat{T}_R$ with SRS–SRS sampling design had smaller errors and was more efficient than $\hat{T}_P$ with PPS–SRS. First-stage PPS designs in conjunction with $\hat{T}_P$ are commonly used in large surveys. This analysis indicated that $\hat{T}_P$ could have serious conditional biases and large errors, and although approximate PPS balance reduced these biases and errors, $\hat{T}_R$ with basket sampling performed better. We found that $\hat{T}_{NL}$ with a BASKET–SRS design out performed all other strategies.

The empirical study used census data that was reasonably described by the model. However, all estimators showed conditional biases. This study and others (Royall and Cumberland 1981a; Cumberland and Royall 1981) have shown that small deviations from the model could cause serious biases. The theoretical results indicated that some conditional biases could be reduced by balanced sampling. The empirical study corroborated these results and showed that first-stage balanced sampling not only reduced con-

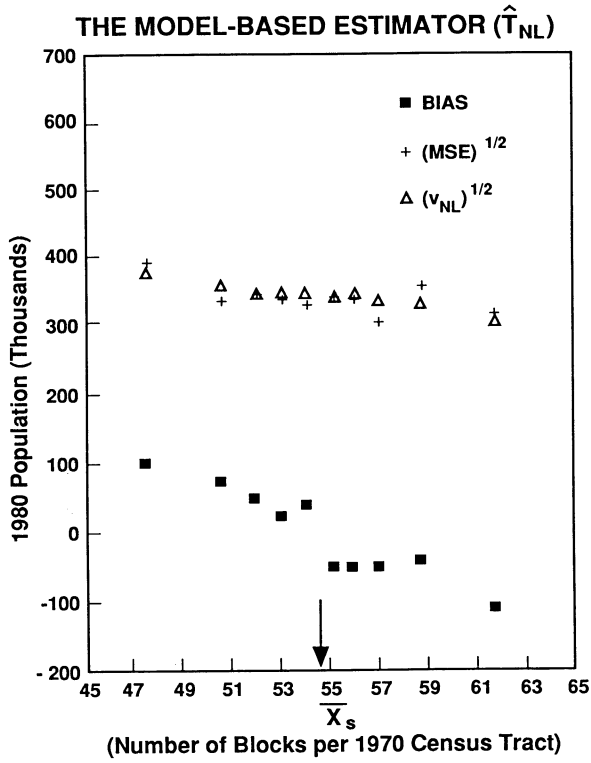

Fig. 6. Bias, $(MSE)^{1/2}$, and $(v_{NL})^{1/2}$ are plotted as functions of the sample means. The bias plot shows the linear dependency of conditional biases on $\bar{X}_s$, which was predicted by the theory under certain types of model failure. The $(MSE)^{1/2}$ and $(v_{NL})^{1/2}$ plots show that the model-based estimator tracks the actual $(MSE)^{1/2}$ quite closely. The arrow marks the balance point.

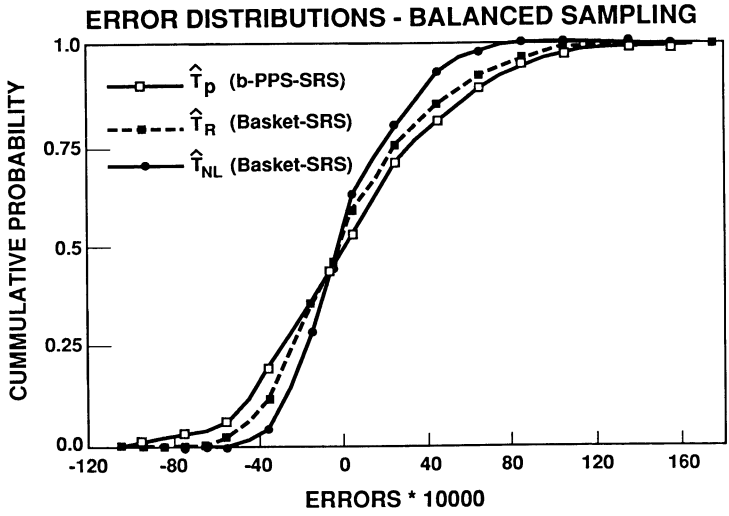

Fig. 7. The cumulative error distributions for $\hat{T}_p$ (b-PPS-SRS), $\hat{T}_R$ (BASKET-SRS), and $\hat{T}_{NL}$ (BASKET-SRS) show that $\hat{T}_{NL}$ (BASKET-SRS) performed best on the real data – less variability and less likely to have large errors.

ditional biases, but also reduced variability and the probability of large errors. Basket sampling produced extremely well-balanced samples (sample moments close to population moments).

The robust variance estimators suggested by Royall for $\hat{T}_P$ and $\hat{T}_R$ tracked the actual MSE quite well. The variance estimator for $\hat{T}_{NL}$ was derived using unbiased sum of squares estimators for unknown quantities in the MSE. Although somewhat cumbersome to compute, this estimator tracked the MSE closely on the real population data.

These theoretical and empirical results indicate that $\hat{T}_{NL}$, its variance estimator, and the BASKET–SRS design deserve further consideration for two-stage samples with unknown cluster size when cluster size is only approximately proportional to a previous size measure and cluster totals are approximate linear functions of cluster size.

8. References

Cochran, W.G. (1977). Sampling Techniques, 3rd ed. New York: John Wiley.

Cohen, B.J. [a.k.a. Kelly, E.J.] (1984). Prediction Theory Approach to Multistage Sampling. Ph.D. Dissertation, University of California at Los Angeles.

Cumberland, W.G. and Royall, M. (1981). Prediction Models and Unequal Probability Sampling. Journal of the Royal Statistical Society, ser. B, 43, 353–367.

Royall, R.M. and Herson, J. (1973). Robust Estimation in Finite Populations. Journal of the American Statistical Association, 68, 880–889.

Royall, R.M. and Cumberland, W.G. (1978). Variance Estimation in Finite Population Sampling. Journal of the American Statistical Association, 73, 351–358.

Royall, R.M. and Cumberland, W.G. (1981a). An Empirical Study of the Ratio Estimator and Estimators of Its Variance. Journal of the American Statistical Association, 76, 66–88.

Royall, R.M. and Cumberland, W.G. (1981b). The Finite Population Linear Regression Estimator and Estimators of Its Variance. Journal of the American Statistical Association, 76, 924–930.

Royall, R.M. (1976). The Least Squares Linear Approach to Two-Stage Sampling. Journal of the American Statistical Association, 71, 657–664.

Royall, R.M. (1986). The Prediction Approach to Robust Variance Estimation in Two-Stage Cluster Sampling. Journal of the American Statistical Association, 81, 119–123.

Rustagi, R.K. (1978). Some Theory of the Prediction Approach to Two-Stage and Stratified Two-Stage Sampling. Ph.D. Dissertation, Ohio State University.

Valliant, R. (1987). Conditional Properties of Some Estimators in Stratified Sampling. Journal of the American Statistical Association, 82, 509–519.

Wallenius, K.J. (1973). On Statistical Methods in Contract Negotiations – Part III. Report N45, Department of Mathematical Sciences, Clemson University.

Received April 1989
Revised June 1990