Predicting random effects with an expanded finite population mixed model

Edward J. Stanek III[a],*, Julio M. Singer[b]

[a]*Department of Public Health, 401 Arnold House, University of Massachusetts, 715 North Pleasant Street, Amherst, MA 01003-9304, USA*
[b]*Departamento de Estatística, Universidade de São Paulo, São Paulo, Brazil*

Abstract

Prediction of random effects is an important problem with expanding applications. In the simplest context, the problem corresponds to prediction of the latent value (the mean) of a realized cluster selected via two-stage sampling. Recently, Stanek and Singer [Predicting random effects from finite population clustered samples with response error. J. Amer. Statist. Assoc. 99, 119–130] developed best linear unbiased predictors (BLUP) under a finite population mixed model that outperform BLUPs from mixed models and superpopulation models. Their setup, however, does not allow for unequally sized clusters. To overcome this drawback, we consider an expanded finite population mixed model based on a larger set of random variables that span a higher dimensional space than those typically applied to such problems. We show that BLUPs for linear combinations of the realized cluster means derived under such a model have considerably smaller mean squared error (MSE) than those obtained from mixed models, superpopulation models, and finite population mixed models. We motivate our general approach by an example developed for two-stage cluster sampling and show that it faithfully captures the stochastic aspects of sampling in the problem. We also consider simulation studies to illustrate the increased accuracy of the BLUP obtained under the expanded finite population mixed model.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Superpopulation; Best linear unbiased predictor; Random permutation; Optimal estimation; Design-based inference; Mixed models

1. Introduction

Optimal estimation of average costs for hospitals that typically vary in size is an important practical problem because of the impact in health care economics, and patient choice of hospital care (see http://www.healthgrades.com, for example). In many cases, this is based on information obtained from patients (units) in hospitals (clusters) realized under a two-stage sampling scheme.

The best linear unbiased predictor (BLUP) developed under a mixed model is often offered as a solution to this problem (Searle et al., 1992). Although the mixed model accounts for unequal numbers of units in sample clusters, it does not use often available information about their sizes. The superpopulation model of Scott and Smith (1969) is an alternative that incorporates this information. Both models can be plausibly used to represent the problem of interest, but neither is formally linked to the finite population from which the two-stage sample is drawn as is the finite

---

* Corresponding author. Tel.: +1 413 545 3812; fax: +1 413 545 1645.

E-mail addresses: stanek@schoolph.umass.edu (E.J. Stanek III), jmsinger@ime.usp.br (J.M. Singer).

Table 1
Population of hospital's appendectomy patients in the past year and observed data

Hospital (s)	M_s	Mean	Variance	Patient* (t)			
s=1 (County)	2	μ_1	σ_1^2	y_{11}	y_{12}		
s=2 (Central)	4	μ_2	σ_2^2	y_{21} $2100 (Jane Blake)	y_{22}	y_{23} $1400 (Sam Evans)	y_{24} $2500 (Hong Yao)
s=3 (Mercy)	2	μ_3	σ_3^2	y_{31} $1700 (Mary Slokum)	y_{32} $1900 (Juan Marcus)	y_{33}	

* Names are fictitious

population mixed model recently proposed by Stanek and Singer (2004)[1] for situations where clusters are of equal size. Under this model, predictors have smaller mean squared error (MSE) than the competitors, even when the variance components are replaced by estimates as indicated in San Martino et al. (2008). We extend the approach of Stanek and Singer (2004) by developing predictors under a new expanded finite population mixed model that outperforms the competitors both in equal and unequal size two-stage cluster sampling problems.

Suppose our interest is in the average cost of appendectomies (the latent value) for each of three hospitals in the past year (Table 1), and that such costs are known (without error) for some patients in two of the hospitals. When the data are obtained from a stratified simple random sample of appendectomy patients, with hospitals as strata, the best linear unbiased estimate is the average cost for the available patients in each hospital (i.e., $2000 for *Central*, and $1800 for *Mercy*).

Now assume that a simple random sample of appendectomy patients is selected from each of a simple random sample of hospitals (Table 2) according to a two-stage sampling scheme. We refer to a sample hospital as a primary sampling unit (PSU) to distinguish it from a specific hospital, and to a sample patient as a secondary sampling unit (SSU) to distinguish it from a specific patient. Under the usual mixed model, the sample appendectomy cost for SSU j in PSU i is

$$Y_{ij} = \mu + B_i + E_{ij}, \tag{1}$$

where μ is the overall mean, B_i is the random effect for PSU i, and E_{ij} is a random variable corresponding to the deviation of the response of SSU j from the latent value of PSU i, namely $T_i = \mu + B_i$. The random variables B_i and E_{ij} are usually considered independent with null expected values, and variances given by σ^2 and σ_i^2, respectively. Model (1) is an example of the general linear mixed model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{B} + \mathbf{E}, \tag{2}$$

where for the sample in Table 2, $\mathbf{X} = \mathbf{1}_r$, $\mathbf{Z} = \oplus_{i=1}^{n}\mathbf{1}_{m_i}$, $\boldsymbol{\alpha} = \mu$, and $\mathbf{B} = (B_1, \ldots, B_n)'$ with $\boldsymbol{\Gamma} = \sigma^2\mathbf{I}_n$, $\boldsymbol{\Sigma} = \oplus_{i=1}^{n}\sigma_i^2\mathbf{I}_{m_i}$, and $\mathrm{var}(\mathbf{Y}) = \boldsymbol{\Omega} = \mathbf{Z}\boldsymbol{\Gamma}\mathbf{Z}' + \boldsymbol{\Sigma}$ with $\mathbf{1}_a$ denoting an $a \times 1$ vector with all elements equal to one, $\mathbf{I}_a$ representing an $a \times a$ identity matrix, and $\oplus_{i=1}^{n}\mathbf{A}_i$ indicating a block diagonal matrix with blocks given by $\mathbf{A}_i$ (Graybill, 1983). This model has a long history (see for example Harville, 1978; Laird and Ware, 1982) and is the main topic in several recent texts such as Brown and Prescott (1999), Verbeke and Molenberghs (2000), McCulloch and Searle (2001), Bryk and Raudenbush (2002), Diggle et al. (2002), Singer and Willett (2003), Demidenko (2004), Littell et al. (2006), and Jiang (2007). Under (1), the BLUP of the latent value for PSU i is

$$\hat{P}_i = \hat{\mu} + k_i(\bar{Y}_i - \hat{\mu}), \tag{3}$$

---

[1] We refer to such models as a finite population mixed models instead of random permutation models as in Stanek and Singer (2004) to avoid confusion with the homonymous, but different, model considered in Hedayat and Sinha (1991).

Table 2
Notation for common mixed model (sample) and superpopulation model (sample and remainder)

	Sample Hospital PSU (i)	Sample Size m_i	Latent Value $\mu + B_i$	Sample Patient SSU (j)			
				j=1	j=2	j=3	j=4
Sample	i=1	3	$\mu + B_1$	Y_{11}	Y_{12}	Y_{13}	
Sample	i=2=n	2	$\mu + B_2$	Y_{21}	Y_{22}		
Remainder	i=1 (Central)						Y_{14}
Remainder	i=2 (Mercy)					Y_{23}	
Remainder	i=3=N (County)			Y_{31}	Y_{32}		

where $\hat{\mu} = \sum_{i=1}^{n}(w_i/\sum_{i=1}^{n}w_i)\bar{Y}_i$ is a weighted sample mean with $w_i = 1/(\sigma^2 + \sigma_i^2/m_i)$, $\bar{Y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i}Y_{ij}$, and $k_i = \sigma^2/(\sigma^2 + \sigma_i^2/m_i)$ (Goldberger, 1962; Henderson, 1984; McLean et al., 1991; Robinson, 1991). The predictor $\hat{P}_i$ is a linear function of $\mathbf{Y}$ (i.e., $\hat{P}_i = \mathbf{L}'\mathbf{Y}$), is unbiased (i.e., $E(\hat{P}_i - T_i) = 0$), and has minimum MSE. Using the realized random variables represented in Table 1, and assuming that $\sigma = 100$, $\sigma_1 = 300$ and $\sigma_2 = 50$, it follows that $\hat{\mu} = \$1844$, $k_1 = 0.25$, $k_2 = 0.89$, and the predictor of the latent value for the realized hospital corresponding to $i = 1$ (i.e., *Central*) is $\hat{P}_1 = \$1883$, while the predictor of the latent value for $i = 2$ (i.e., *Mercy*) is $\hat{P}_2 = \$1805$.

The estimate of a realized hospital's latent value derived from the stratified model or the corresponding predictor obtained from the mixed model do not use additional information, such as the number of hospitals in the population, or the number of appendectomy patients in each hospital, even though such additional information may be available (as illustrated in the remainder in Table 2). The combined sample and remainder represents a superpopulation that is constructed by first (conceptually) selecting a finite population (presumably from some larger population in time or space), and then selecting a two-stage sample from it. Scott and Smith (1969) show that the latent value for a hospital in the superpopulation, $T_i = \sum_{j=1}^{M_i}Y_{ij}/M_i$, is predicted by

$$\hat{P}_i = f_i\bar{Y}_i + (1 - f_i)[\hat{\mu} + k_i(\bar{Y}_i - \hat{\mu})], \tag{4}$$

where $f_i = m_i/M_i$. Using the data in Table 1, the resulting predictor for $i = 1$ (i.e., *Central*) is $\hat{P}_1 = \$1971$, and for $i = 2$ (i.e., *Mercy*) is $\hat{P}_2 = \$1802$.

The superpopulation model does not clearly separate the labeled clusters (as in Table 1) from random variables that represent a sample of clusters (i.e., note how *Central* Hospital is uniquely associated with $i = 1$ in Table 2). This separation is clear when the two-stage sampling process is represented with indicator random variables in the finite population mixed model developed by Stanek and Singer (2004). The resulting predictor (limited to equal size clusters and equal size cluster sample sizes) is

$$\hat{P}_i = f\bar{Y}_i + (1 - f)[\bar{Y} + k(\bar{Y}_i - \bar{Y})], \tag{5}$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i$, $f = \frac{m}{M}$, $k = \sigma^{*2}/(\sigma^{*2} + \sigma_e^2/m)$, $\sigma^{*2} = \sigma^2 - \frac{\sigma_e^2}{M}$ and $\sigma_e^2 = \frac{1}{N}\sum_{s=1}^{N}\sigma_s^2$. With equal size clusters, the predictor specified by (4) differs from (5) since variance components have different definitions. Theoretically, the expected MSE of (5) is less than the expected MSE for (4) or (3) as shown by Stanek and Singer (2004), while the empirical version of (5), formed by replacing variance components with their sample estimates, in general outperforms the empirical versions of the other predictors, as summarized by San Martino et al. (2008). However, (5) cannot be used for data like those in Tables 1 and 2 since cluster and sample sizes differ.

When clusters are of equal size, the finite population mixed model can be used to represent the remaining random variables (as in Table 2) without the need to identify the realized clusters for sample PSUs. When clusters differ in size, we do not know how many SSUs remain since we do not know the size of the realized PSU. Other problems occur with the representation in Table 2, as for example, the impossibility that the PSU $i = 1$ be *County* Hospital (Table 2), even though the first stage sampling is assumed to be simple random sampling, or the apparent random nature of the second stage sample size, PSU size, and SSU variance due to the first stage sampling.

We extend the expanded model used by Stanek et al. (2004) for simple random sampling to two-stage unbalanced sampling to overcome these problems. The expanded model simultaneously retains the cluster identity and the PSU position, and for each PSU, distinguishes the relevant contribution of both sampled SSUs, and non-sampled SSUs to a target random variable such as a PSU mean. For such purposes, we first define an expanded set of random variables, and subsequently show that a lower dimensional (collapsed) set can adequately represent the problem without loss of information. Following the steps in Stanek and Singer (2004), we specify the expanded finite population mixed model in Section 2, derive the corresponding BLUP along with its theoretical expected MSE in Section 3, and compare the proposed predictor to others via simulation studies in Section 4, and conclude with discussion in Section 5.

2. An expanded mixed model for a finite clustered population

Let a finite population be defined (as in Table 1) by a listing of units, labeled by $t = 1, \ldots, M_s$, in each cluster, labeled by $s = 1, \ldots, N$, where the non-stochastic potentially observable response for unit t in cluster s is given by y_{st}. The finite population mean and variance for cluster s are respectively defined as $\mu_s = \frac{1}{M_s} \sum_{t=1}^{M_s} y_{st}$ and $((M_s - 1)/M_s)\sigma_s^2 = \frac{1}{M_s} \sum_{t=1}^{M_s} (y_{st} - \mu_s)^2$. Similarly, the population mean, and between cluster variance are respectively defined as $\mu = \frac{1}{N} \sum_{s=1}^{N} \mu_s$ and $((N - 1)/N)\sigma^2 = \frac{1}{N} \sum_{s=1}^{N} (\mu_s - \mu)^2$. We represent the potentially observable response for unit t in cluster s as $y_{st} = \mu + \beta_s + \varepsilon_{st}$ where $\beta_s = (\mu_s - \mu)$ is the deviation of the mean for cluster s from the overall mean, and $\varepsilon_{st} = (y_{st} - \mu_s)$ is the deviation of the response for unit t (in cluster s) from the mean for cluster s. Letting $\mathbf{y} = (\mathbf{y}_1' \ \mathbf{y}_2' \ \cdots \ \mathbf{y}_N')'$ where $\mathbf{y}_s = (y_{s1} \ y_{s2} \ \cdots \ y_{sM_s})'$, the reparameterized finite population can be summarized as

$$\mathbf{y} = \mathbf{X}\mu + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6}$$

where $\mathbf{X} = \mathbf{1}_{\mathbb{N}}$, $\mathbf{Z}_{\mathbb{N} \times N} = \oplus_{s=1}^{N} \mathbf{1}_{M_s}$, $\mathbb{N} = \sum_{s=1}^{N} M_s$, $\boldsymbol{\beta} = (\beta_1 \ \beta_2 \ \cdots \ \beta_N)'$, and $\boldsymbol{\varepsilon}$ is defined similarly to $\mathbf{y}$. None of the terms in (6) are random variables.

2.1. The expanded set of random variables

We define a vector of random variables to represent equally likely two-stage random permutations of the population (i.e., with probability $1/(N! \prod_{s=1}^{N} M_s!)$, as in Cochran, 1977). Without loss of generality, we assume that the sample clusters are in the first n positions in a permutation of clusters and that the sample units in cluster s correspond to the units in the first m_s positions in a permutation of that cluster's units. The ordering of the random variables is important, since any realization can be re-ordered to exactly match the finite population values.

We use indicator random variables to relate the response for unit t in cluster s, namely, y_{st}, to the response for SSU j in PSU i. To do so, we let $U_{jt}^{(s)}$ be an indicator random variable that takes on a value of one when SSU j in cluster s is unit t, and zero otherwise, so that the response for SSU j in cluster s may be expressed as $\tilde{Y}_{sj} = \sum_{t=1}^{M_s} U_{jt}^{(s)} y_{st}$. We include a fixed non-stochastic weight w_{sj} for SSU j in cluster s, and define the weighted response as $\tilde{Y}_{wsj} = w_{sj} \tilde{Y}_{sj}$ so that the sum, $\sum_{j=1}^{M_s} \tilde{Y}_{wsj}$, will correspond to a cluster total when $w_{sj} = 1$ for all $j = 1, \ldots, M_s$, or to a cluster mean when $w_{sj} = \frac{1}{M_s}$ for all $j = 1, \ldots, M_s$, for example. Letting $\mathbf{U}_j^{(s)} = (U_{j1}^{(s)} \ U_{j2}^{(s)} \ \cdots \ U_{jM_s}^{(s)})'$, it follows that $\tilde{Y}_{wsj} = w_{sj} \mathbf{y}_s' \mathbf{U}_j^{(s)}$. The vector $\tilde{\mathbf{Y}}_{ws} = (\tilde{Y}_{ws1} \ \tilde{Y}_{ws2} \ \cdots \ \tilde{Y}_{wsM_s})'$ represents a permutation of weighted responses for the SSUs in cluster s.

We also let U_{is} be an indicator random variable that takes on a value of one when PSU i is cluster s, and a value of zero otherwise. If all clusters were equal in size, we could represent a permutation of SSUs for PSU i by $\sum_{s=1}^{N} U_{is} \tilde{\mathbf{Y}}_{ws}$. When cluster sizes differ, this sum is not defined, since the dimensions of the vectors composing it cannot all be

equal. We solve this problem by expanding the set of random variables associated to PSU i into the $\mathbb{N} \times 1$ vector $\overset{\leftrightarrow}{\mathbf{Y}}_{wi} = ((U_{is}\tilde{\mathbf{Y}}_{ws})) = (U_{i1}\tilde{\mathbf{Y}}'_{w1} \ U_{i2}\tilde{\mathbf{Y}}'_{w2} \ \cdots \ U_{iN}\tilde{\mathbf{Y}}'_{wN})'$ so that a two-stage random permutation of the population is then represented by the $N\mathbb{N} \times 1$ vector, $\overset{\leftrightarrow}{\mathbf{Y}}_w = ((\overset{\leftrightarrow}{\mathbf{Y}}_{wi})) = (\overset{\leftrightarrow}{\mathbf{Y}}'_{w1} \ \overset{\leftrightarrow}{\mathbf{Y}}'_{w2} \ \cdots \ \overset{\leftrightarrow}{\mathbf{Y}}'_{wN})'$, where the jth element of $\overset{\leftrightarrow}{\mathbf{Y}}_w$ that corresponds to U_{is} is $U_{is}\tilde{Y}_{wsj}$.

2.2. The expanded finite population mixed model

We construct a mixed model for the expanded response vector $\overset{\leftrightarrow}{\mathbf{Y}}_w$ next. Indexing expectation with respect to permutations of clusters with the subscript ξ_1 and expectation with respect to permutations of units in a cluster with the subscript ξ_2, and for PSU i, we let

$$\overset{\leftrightarrow}{\mathbf{Y}}_{wi} = E_{\xi_1\xi_1}(\overset{\leftrightarrow}{\mathbf{Y}}_{wi}) + [E_{\xi_2|\xi_1}(\overset{\leftrightarrow}{\mathbf{Y}}_{wi}) - E_{\xi_1\xi_1}(\overset{\leftrightarrow}{\mathbf{Y}}_{wi})] + \overset{\leftrightarrow}{\mathbf{E}}_{wi},$$

where $E_{\xi_1\xi_2}(\overset{\leftrightarrow}{\mathbf{Y}}_{wi}) = \frac{1}{N}(\oplus_{s=1}^N \mathbf{w}_s)\boldsymbol{\mu}$, $E_{\xi_2|\xi_1}(\overset{\leftrightarrow}{\mathbf{Y}}_{wi}) = (\oplus_{s=1}^N \mathbf{w}_s\mu_s)\mathbf{U}_i$, $\overset{\leftrightarrow}{\mathbf{E}}_{wi} = \overset{\leftrightarrow}{\mathbf{Y}}_{wi} - E_{\xi_2|\xi_1}(\overset{\leftrightarrow}{\mathbf{Y}}_{wi})$,

$$\mathbf{w}_s = ((w_{sj})) = (w_{s1} \ w_{s2} \ \cdots \ w_{sM_s})', \quad \boldsymbol{\mu} = ((\mu_s)) = (\mu_1 \ \mu_2 \ \cdots \ \mu_N)',$$

$\mathbf{U}_i = ((U_{is})) = (U_{i1} \ U_{i2} \ \cdots \ U_{iN})'$ and $\overset{\leftrightarrow}{\mathbf{E}}_{wi}$ denotes the deviation of response from the expected response within a PSU. The fixed effects are given by $\boldsymbol{\mu}$, the vector of cluster means, while the random effects correspond to $E_{\xi_2|\xi_1}(\overset{\leftrightarrow}{\mathbf{Y}}_{wi}) - E_{\xi_1\xi_1}(\overset{\leftrightarrow}{\mathbf{Y}}_{wi})$. In the finite population mixed model of Stanek and Singer (2004), the random effect for PSU i was defined as $\sum_{s=1}^N U_{is}\beta_s = \sum_{s=1}^N(U_{is}\mu_s - U_{is}\sum_{s^*=1}^N \frac{1}{N}\mu_{s^*})$, with the random variables U_{is} explicitly linking the clusters to PSU i. In the finite expanded mixed model, random effects are defined for SSU j in PSU i as $w_{sj}\mu_s(U_{is} - E_{\xi_1}(U_{is}))$. For example, when $w_{sj} = \frac{1}{M_s}$ for all $j = 1, \ldots, M_s$ (corresponding to the PSU mean), it follows that $\sum_{j=1}^{M_s} w_{sj}\mu_s(U_{is} - E_{\xi_1}(U_{is})) = U_{is}\mu_s - \frac{1}{N}\mu_s$. For both models, the expected value of the random effects (with respect to ξ_1) is zero. We combine the fixed and random effects to define the expanded finite population mixed model as

$$\overset{\leftrightarrow}{\mathbf{Y}}_w = \left[\frac{1}{N}\mathbf{1}_N \otimes \left(\overset{N}{\underset{s=1}{\oplus}} \mathbf{w}_s\right)\right]\boldsymbol{\mu} + \left[\mathbf{I}_N \otimes \left(\overset{N}{\underset{s=1}{\oplus}} \mathbf{w}_s\mu_s\right)\right]\text{vec}(\mathbf{U} - E_{\xi_1}(\mathbf{U})) + \overset{\leftrightarrow}{\mathbf{E}}_w \tag{7}$$

where $\mathbf{U} = (\mathbf{U}_1 \ \mathbf{U}_2 \ \cdots \ \mathbf{U}_N)$. The covariance matrix of the random effects is $\text{var}_{\xi_1\xi_2}([\mathbf{I}_N \otimes (\oplus_{s=1}^N \mathbf{w}_s\mu_s)]\text{vec}(\mathbf{U} - E_{\xi_1}(\mathbf{U}))) = \frac{1}{N-1}\mathbf{P}_N \otimes [(\oplus_{s=1}^N \mathbf{w}_s\mu_s)\mathbf{P}_N(\oplus_{s=1}^N \mathbf{w}_s\mu_s)]$ while the covariance matrix of $\overset{\leftrightarrow}{\mathbf{E}}_w$ is

$$\text{var}_{\xi_1\xi_2}(\overset{\leftrightarrow}{\mathbf{E}}_w) = \mathbf{I}_N \otimes \left(\overset{N}{\underset{s=1}{\oplus}}\left[\frac{\sigma_s^2}{N}\left(\overset{M_s}{\underset{j=1}{\oplus}} w_{sj}\right)\mathbf{P}_{M_s}\left(\overset{M_s}{\underset{j=1}{\oplus}} w_{sj}\right)\right]\right),$$

where $\mathbf{P}_a = \mathbf{I}_a - \frac{1}{a}\mathbf{J}_a$ and $\mathbf{J}_a$ denotes an $a \times a$ matrix with all elements equal to one.

2.3. Defining target quantities

Model (7) is an expanded version of a finite population mixed model that retains the identity of clusters, while accounting for a two-stage random permutation. Our interest is to predict target linear combinations defined by $T = \mathbf{g}'\overset{\leftrightarrow}{\mathbf{Y}}_w$, where $\mathbf{g}$ is non-stochastic. For simplicity, we limit discussion to the case

$$\mathbf{g}' = \mathbf{c} \otimes \mathbf{1}'_{\mathbb{N}} \tag{8}$$

where $\mathbf{c}$ is an $N \times 1$ vector of constants. In particular, we focus on the setting where $\mathbf{c} = \mathbf{e}_i$, i.e., an $N \times 1$ vector with all elements equal to zero, except for element i which has the value of one. The principal interest lies in the setting where $i \leqslant n$, i.e., in the clusters realized in the sample. When $w_{sj} = \frac{1}{M_s}$ for all $s = 1, \ldots, N, j = 1, \ldots, M_s$, the target, $T = \sum_{s=1}^N U_{is}(\sum_{j=1}^{M_s} w_{sj}\tilde{Y}_{sj})$ is the mean of PSU i; when $w_{sj} = 1$ for all $s = 1, \ldots, N, j = 1, \ldots, M_s$, the target, $T = \sum_{s=1}^N U_{is}(\sum_{j=1}^{M_s} w_{sj}\tilde{Y}_{sj})$ is the total of PSU i. Note that in both cases, the target is a random variable.

3. Predicting a PSU mean in the expanded finite population mixed model

To obtain the BLUP, we adopt the basic strategy employed by Scott and Smith (1969), Royall (1976), Bolfarine and Zacks (1992), Valliant et al. (2000), and Stanek and Singer (2004), among others. We assume that the elements in the sample portion of $\overset{\leftrightarrow}{\mathbf{Y}}_w$ will be observed, and express the target T as the sum of two parts, one which is a function of the sample, and the other, a function of the remaining random variables. Then, requiring the predictor to be a linear function of the sample random variables and to be unbiased, we obtain coefficients that minimize the MSE. While in theory, an optimal predictor can be obtained via this recipe, in practice, the high dimensionality of the expanded random vectors may result in singularities that lead to multiple solutions as discussed in Stanek et al. (2004). For this reason, we explore projections of the expanded random variables into lower dimensional spaces that retain the necessary information for an optimal solution.

3.1. Partial collapsing of the expanded finite population mixed model random variables

Following Rao and Bellhouse (1978), we provide a way of determining whether the optimal linear unbiased predictor of a target random variable, $T = \mathbf{g}'\overset{\leftrightarrow}{\mathbf{Y}}_w$ can be obtained as the optimal linear unbiased predictor of $T = \mathbf{g}'_p \overset{\leftrightarrow}{\mathbf{Y}}_{wp}$ based on a vector of collapsed random variables that spans a lower dimensional space defined by $\overset{\leftrightarrow}{\mathbf{Y}}_{wp} = \overset{\leftrightarrow}{\mathbf{C}}{}' \overset{\leftrightarrow}{\mathbf{Y}}_w$, where $\overset{\leftrightarrow}{\mathbf{C}}$ is a matrix of dimension $N\mathbb{N} \times c$ with $c < N\mathbb{N}$. We take

$$
\ddot{\mathbf{C}}' = \begin{pmatrix} \overset{N}{\underset{i=1}{\oplus}} \overset{N}{\underset{s=1}{\oplus}} \left(\mathbf{1}'_{m_s} \ \vdots \ \mathbf{0}'_{M_s - m_s} \right) \\ \hline \overset{N}{\underset{i=1}{\oplus}} \overset{N}{\underset{s=1}{\oplus}} \left(\mathbf{0}'_{m_s} \ \vdots \ \mathbf{1}'_{M_s - m_s} \right) \end{pmatrix}
$$

and $\mathbf{g}'_p = \mathbf{g}'[\overset{\leftrightarrow}{\mathbf{C}}(\overset{\leftrightarrow}{\mathbf{C}}{}' \overset{\leftrightarrow}{\mathbf{C}})^{-1}]$, so that the effect of collapsing is the generation of sums of the SSUs for the sample and for the remainder in each cluster for each PSU, thus reducing the number of random variables from $N\mathbb{N}$ to $2N^2$. Since $\overset{\leftrightarrow}{\mathbf{Y}}_w = [\overset{\leftrightarrow}{\mathbf{C}}(\overset{\leftrightarrow}{\mathbf{C}}{}'\overset{\leftrightarrow}{\mathbf{C}})^{-1}]\overset{\leftrightarrow}{\mathbf{Y}}_{wp} + \mathbf{P}_{\overset{\leftrightarrow}{\mathbf{C}}} \overset{\leftrightarrow}{\mathbf{Y}}_w$, where $\mathbf{P}_{\overset{\leftrightarrow}{\mathbf{C}}} = \mathbf{I}_{N\mathbb{N}} - \overset{\leftrightarrow}{\mathbf{C}}(\overset{\leftrightarrow}{\mathbf{C}}{}'\overset{\leftrightarrow}{\mathbf{C}})^{-1}\overset{\leftrightarrow}{\mathbf{C}}{}'$, we can write $\mathbf{g}'\overset{\leftrightarrow}{\mathbf{Y}}_w = \mathbf{g}'[\overset{\leftrightarrow}{\mathbf{C}}(\overset{\leftrightarrow}{\mathbf{C}}{}'\overset{\leftrightarrow}{\mathbf{C}})^{-1}]\overset{\leftrightarrow}{\mathbf{Y}}_{wp} + \mathbf{g}'\mathbf{P}_{\overset{\leftrightarrow}{\mathbf{C}}}\overset{\leftrightarrow}{\mathbf{Y}}_w$. Using (8), it follows that $\mathbf{g}'_p = \mathbf{1}'_2 \otimes [\mathbf{c}' \otimes \mathbf{1}'_N]$, $\mathbf{g}'\mathbf{P}_{\overset{\leftrightarrow}{\mathbf{C}}}\overset{\leftrightarrow}{\mathbf{Y}}_w = 0$ and $T = \mathbf{g}'_p\overset{\leftrightarrow}{\mathbf{Y}}_{wp}$. Let $\hat{\mathbf{L}}_p$ represent an $nN \times 1$ vector of constants, and $\overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{I}}$ be the first nN random variables (corresponding to the sample) in $\overset{\leftrightarrow}{\mathbf{Y}}_{wp}$. Then letting $\hat{T}_p = \hat{\mathbf{L}}'_p \overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{I}}$ be the optimal linear unbiased predictor of T based on $\overset{\leftrightarrow}{\mathbf{Y}}_{wp}$, and $\hat{B}_p$ be a linear unbiased predictor of $\mathbf{g}'\mathbf{P}_{\overset{\leftrightarrow}{\mathbf{C}}}\overset{\leftrightarrow}{\mathbf{Y}}_w = 0$, it follows from Rao and Bellhouse (1978, Theorem 1.1) that $\hat{T}_p$ will be optimal for $T = \mathbf{g}'\overset{\leftrightarrow}{\mathbf{Y}}_w$ if and only if $E_{\xi_1 \xi_2}[(\hat{T}_p - T)\hat{B}_p] = 0$. Expressing $E_{\xi_1 \xi_2}[(\hat{T}_p - T)\hat{B}]$ as a function of $E_{\xi_1 \xi_2}(\overset{\leftrightarrow}{\mathbf{Y}}_w \overset{\leftrightarrow}{\mathbf{Y}}'_w)$, and simplifying terms, it follows that $E_{\xi_1 \xi_2}[(\hat{T}_p - T)\hat{B}_p] = 0$ when $w_{sj} = w_s$ for all $j = 1, \ldots, M_s$ (see details at http://www.umass.edu/cluster/). This implies that we can obtain the optimal predictor using the partially collapsed random variables as long as within each cluster, the weights are equal for all SSUs.

Having this in mind, we assume that $w_{sj} = w_s$ for all $j = 1, \ldots, M_s$ and develop the BLUP of $T = \mathbf{g}'_p\overset{\leftrightarrow}{\mathbf{Y}}_{wp}$ based on the $2N^2$ collapsed random variables contained in $\overset{\leftrightarrow}{\mathbf{Y}}_{wp}$. The first N^2 of them are of the form $U_{is} w_s m_s \bar{Y}_{s\mathrm{I}}$, while the remaining N^2 are of the form $U_{is} w_s (M_s - m_s)\bar{Y}_{s\mathrm{II}}$, where $\bar{Y}_{s\mathrm{I}} = \frac{1}{m_s}\sum_{j=1}^{m_s}\tilde{Y}_{sj}$ and $\bar{Y}_{s\mathrm{II}} = \frac{1}{M_s - m_s}\sum_{j=m_s+1}^{M_s}\tilde{Y}_{sj}$.

3.2. Predicting linear combinations of PSU latent values using the expanded finite population mixed model

We partition $\overset{\leftrightarrow}{\mathbf{Y}}_{wp}$ into the first nN random variables corresponding to the sample, $\overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{I}}$, and the $N(N - n)$ remaining random variables, $\overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{II}}$, and write the target as $T = \mathbf{g}'_{\mathrm{I}}\overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{I}} + \mathbf{g}'_{\mathrm{II}}\overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{II}}$, where $\mathbf{g}'_{\mathrm{I}} = \mathbf{c}'_{\mathrm{I}} \otimes \mathbf{1}'_N$ and $\mathbf{g}'_{\mathrm{II}} = \left(\mathbf{c}'_{\mathrm{II}} \otimes \mathbf{1}'_N \ \vdots \ \mathbf{c}' \otimes \mathbf{1}'_N \right)$.

Explicitly, the partitioned expanded finite population mixed model is

$$\begin{pmatrix} \overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{I}} \\ \overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{II}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{\mathrm{I}} \\ \mathbf{X}_{\mathrm{II}} \end{pmatrix} \boldsymbol{\mu} + \left[E_{\xi_2} \begin{pmatrix} \overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{I}} \\ \overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{II}} \end{pmatrix} - E_{\xi_1 \xi_2} \begin{pmatrix} \overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{I}} \\ \overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{II}} \end{pmatrix} \right] + \begin{pmatrix} \overset{\leftrightarrow}{\mathbf{E}}_{w\mathrm{I}} \\ \overset{\leftrightarrow}{\mathbf{E}}_{w\mathrm{II}} \end{pmatrix}. \tag{9}$$

Requiring the predictor of T to be a linear function of $\overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{I}}$ to be unbiased, and to have minimum MSE, the BLUP of T in (9) is

$$\hat{T}_p = \sum_{i=1}^{n} c_i (\hat{Y}_i - \bar{\hat{Y}}) + \bar{c} \frac{N}{n} \left(\sum_{s=1}^{N} I_s (M_s w_s \bar{Y}_{s\mathrm{I}}) \right) + \bar{c}_{\mathrm{II}} \frac{N-n}{n} \left(\sum_{s=1}^{N} I_s (M_s f_s w_s \bar{Y}_{s\mathrm{I}}) \right), \tag{10}$$

where $\hat{Y}_i = \sum_{s=1}^{N} U_{is} M_s w_s k_s^* \bar{Y}_{s\mathrm{I}}$, $\bar{\hat{Y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i$, $k_s^* = k_s - \frac{k_s}{d_s} \frac{1}{N} \sum_{s^*=1}^{N} (\frac{1-k_{s^*}}{1-\bar{k}}) d_{s^*}$, $d_s = M_s w_s \mu_s$, $k_s = \frac{f_s^2 d_s^2}{f_s^2 d_s^2 + (N-1) v_{se}^2}$ (see a: http://www.umass.edu/cluster/ for details), $\bar{k} = \frac{1}{N} \sum_{s=1}^{N} k_s$, $\bar{c}_{\mathrm{II}} = \frac{1}{N-n} \sum_{i=n+1}^{N} c_i$, $v_{se}^2 = f_s (1 - f_s) \frac{M_s w_s^2 \sigma_s^2}{N}$, $\bar{c} = \frac{1}{N} \sum_{i=1}^{N} c_i$ and $I_s = \sum_{i=1}^{n} U_{is}$ is an indicator 'inclusion' random variable for cluster s in the sample (see derivation in Appendix A). An expression for the MSE of the predictor can be developed directly using expressions for the variance, and simplifies to

$$\mathrm{var}_{\xi_1 \xi_2}(\hat{T}_p - T) = \left(\sum_{i=1}^{n} (c_i - \bar{c}_{\mathrm{I}})^2 \right) \left(\sigma_{kd}^2 - 2\sigma_{kd,d} + \frac{1}{N} \sum_{s=1}^{N} \frac{k_s^{*2} v_{se}^2}{f_s^2} \right) + \frac{(N\bar{c})^2}{n} \left(\frac{1}{N} \sum_{s=1}^{N} \frac{v_{se}^2}{f_s^2} \right)$$

$$+ \left[\sum_{i=1}^{N} c_i^2 + \frac{N\bar{c}}{n} (N\bar{c} - 2n\bar{c}_{\mathrm{I}}) \right] \sigma_d^2,$$

where $\bar{c}_{\mathrm{I}} = \frac{1}{n} \sum_{i=1}^{n} c_i$, $\mu_d = \frac{1}{N} \sum_{s=1}^{N} d_s$, $\mu_{kd} = \frac{1}{N} \sum_{s=1}^{N} k_s^* d_s$, $\sigma_{kd}^2 = \frac{1}{N-1} \sum_{s=1}^{N} (k_s^* d_s - \mu_{kd})^2$, $\sigma_d^2 = \frac{1}{N-1} \sum_{s=1}^{N} (d_s - \mu_d)^2$ and $\sigma_{kd,d} = \frac{1}{N-1} \sum_{s=1}^{N} (k_s^* d_s - \mu_{kd})(d_s - \mu_d)$ (see b. http://www.umass.edu/cluster/ for details).

When predicting a PSU mean, i.e., using $w_s = \frac{1}{M_s}$, (10) simplifies to $\hat{T}_p = \bar{Y} + (\hat{Y}_i - \bar{\hat{Y}})$ if $i \leqslant n$, and to $\hat{T}_p = \bar{Y}$ when $i > n$, where $\bar{Y}_i = \sum_{s=1}^{N} U_{is} \bar{Y}_{s\mathrm{I}}$. The MSE for a sample PSU mean predictor (when $i \leqslant n$) simplifies to

$$\mathrm{var}_{\xi_1 \xi_2}(\hat{T}_p - T) = \left(\frac{n-1}{n} \right) \frac{1}{N-1} \sum_{s=1}^{N} \left[\left((k_s^* - 1)\mu_s - \frac{1}{N} \sum_{s=1}^{N} (k_s^* - 1)\mu_s \right)^2 \right]$$

$$+ \frac{1}{nN} \sum_{s=1}^{N} \left[1 + (n-1)k_s^{*2} \right] (1 - f_s) \frac{\sigma_s^2}{m_s}$$

(see c. http://www.umass.edu/cluster/ for details) while the MSE for a PSU not in the sample is given by

$$\mathrm{var}_{\xi_1 \xi_2}(\hat{T}_p - T) = \left(\frac{n+1}{n} \right) \sigma^2 + \frac{1}{nN} \sum_{s=1}^{N} (1 - f_s) \frac{\sigma_s^2}{m_s}.$$

4. Comparison of predictors

We compare the MSE of (10) to that of the simple mean, and of predictors (3) and (4). When clusters are of equal size, have homogeneous unit variances, and sample sizes are equal, the MSE for each predictor can be explicitly calculated. In this setting, we also compare the results with the MSE for predictor (5). For the sample mean,

$$\mathrm{MSE}(\bar{Y}_i) = \frac{1}{N} \sum_{s=1}^{N} (1 - f_s) \frac{\sigma_s^2}{m_s}, \text{ while for (5), } \mathrm{MSE}(\hat{T}_{RP}) = (1 - f) \left[\frac{\sigma_e^2}{nm} + \left(\frac{n-1}{n} \right) (1 - k)\sigma^2 \right].$$
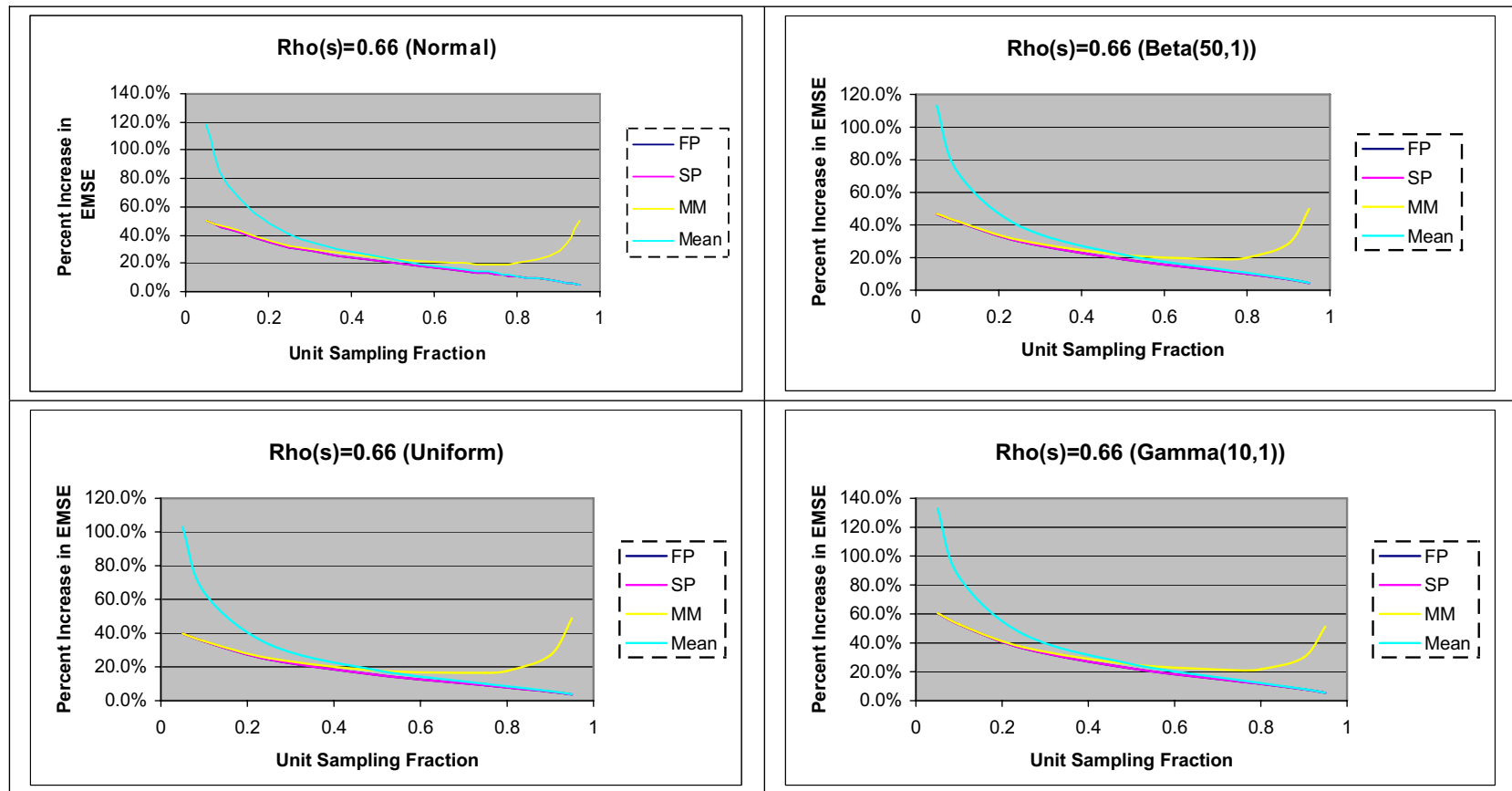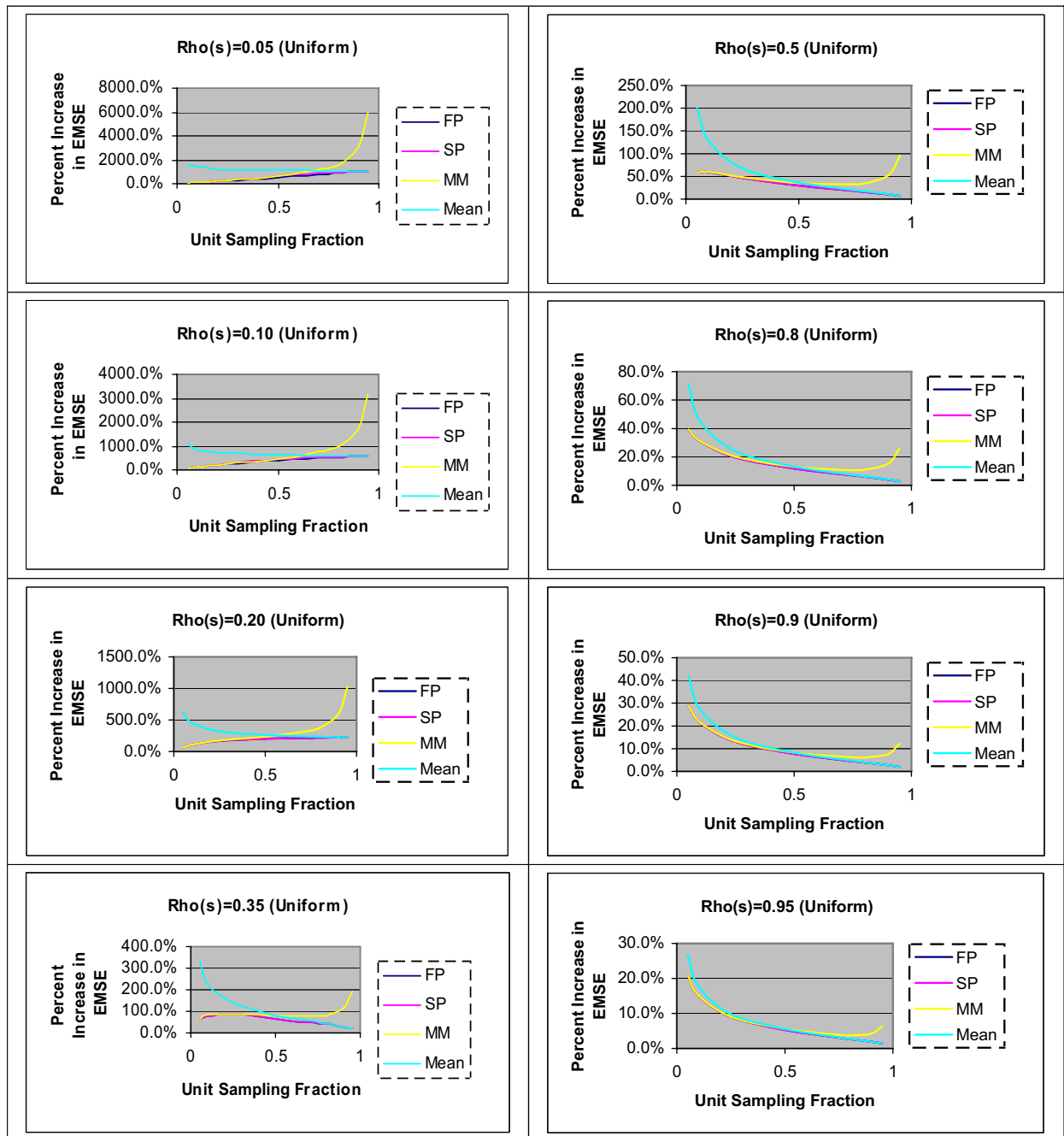
Fig. 1. Percent increase in MSE for the finite population mixed model (FP), superpopulation model (SP), mixed model (MM), and sample mean (Mean) predictors relative to finite expanded mixed model predictor of a realized PSU mean where $N = 100$, $n = 30$, and $M = 20$ for all clusters. Equal size clusters and equal unit sampling fractions per cluster. *Source*: ced07p34.xls.

Fig. 2. Percent increase in MSE for the finite population mixed model (FP), superpopulation model (SP), mixed model (MM), and sample mean (mean) predictors relative to finite expanded mixed model predictor of a realized PSU mean by unit intra-class correlation and unit sampling fraction where $N = 100$, $n = 30$, and $M = 20$ for all clusters. Equal size clusters and equal unit sampling fractions per cluster. *Source*: ced07p38.sas.

The MSE for predictors (3) and (4) are given by $\mathrm{MSE}(\hat{T}_{RP}) = ((n-1)/nk)(\sigma^2 - \sigma_e^2/M)(c - [f + (1-f)k])^2$ with $c = m\sigma^2/(m\sigma^2 + \sigma_e^2)$ for (3) and $c = f + (1-f)m\sigma^2/(m\sigma^2 + \sigma_e^2)$ for (4) as shown in Stanek and Singer (2004). Although we have explicit expressions for the MSE of these predictors, the difference between them is a complicated function of the population parameters. Since shrinkage constants for the expanded predictor depend on the cluster latent values, we compare the MSE relative to the expanded finite population mixed model predictor in four settings with
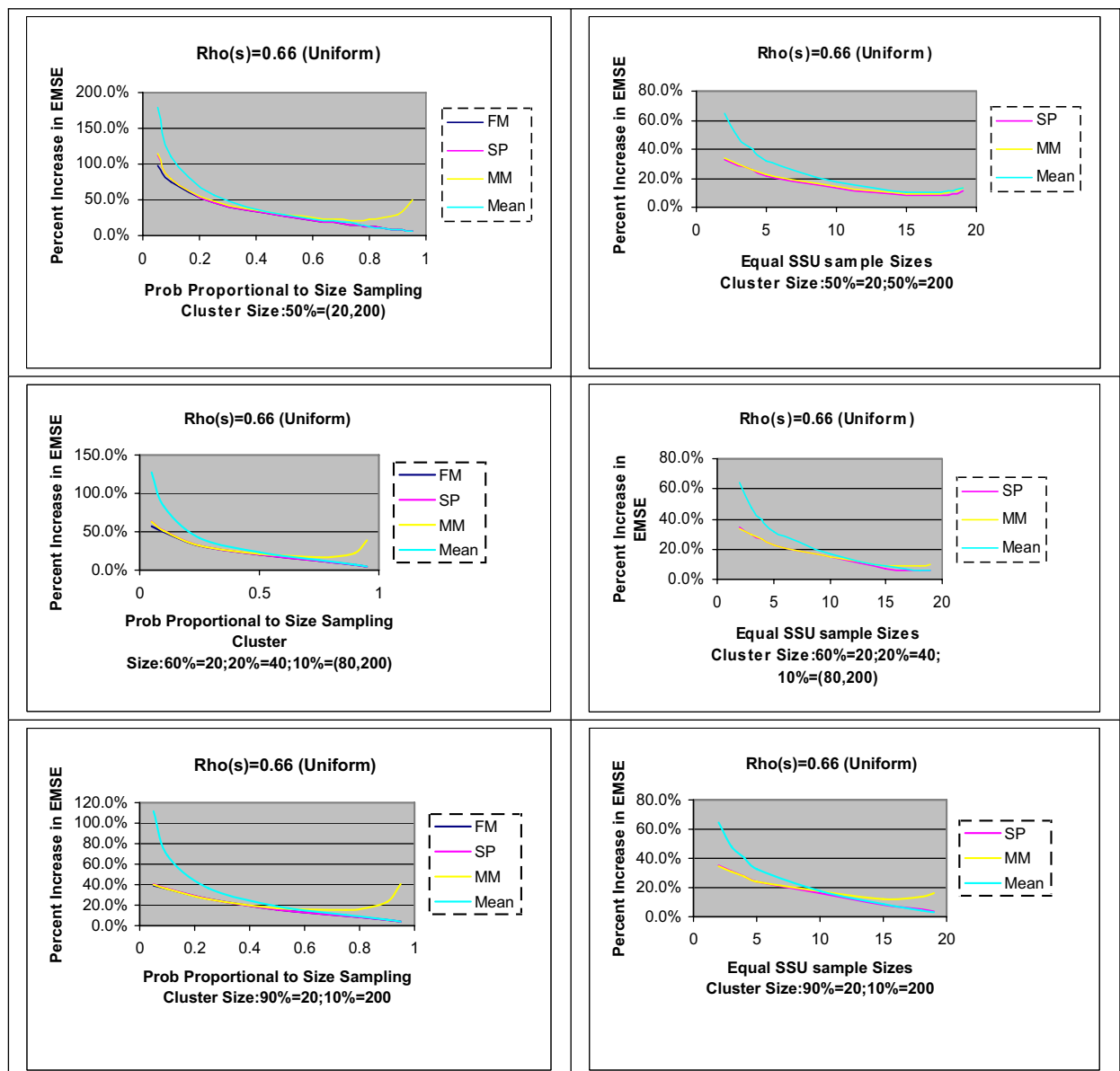
Fig. 3. Percent increase in MSE for the finite population mixed model (FM), superpopulation model (SP), mixed model (MM), and sample mean (mean) predictors relative to finite expanded mixed model predictor of a realized PSU mean with probability proportional to size SSU sampling and for equal SSU size sampling for different cluster sizes where $N = 100$, $n = 30$, and $M = 20$ for all clusters. *Source*: ced07p41.xls.

different values of the unit intra-class correlation coefficient, $\rho_s = \sigma^2/(\sigma^2 + \sigma_s^2)$. In each setting, the cluster latent values are set equal to evenly spaced quantiles from some specified distribution. The results, expressed as percent increase in MSE relative to the MSE of (10) are presented in Fig. 1, and illustrate that in all settings considered, using (10) results in a substantial reduction in MSE (over 40% when $f < 0.2$). This is true even for (5), illustrating that a smaller MSE can be achieved for the BLUP derived under the expanded finite population mixed model as compared to the BLUP obtained under the finite population mixed model of Stanek and Singer (2004). There were little differences in the MSE comparisons with different distributions of the cluster latent values. The results illustrate that predictor (3) has larger MSE relative to the other predictors when $f > 0.5$. The MSE for predictors (4) and (5) are similar, and differ more from the MSE of (10) when f is small.

Fig. 2 summarizes increases in MSE for different intra-class correlation coefficients; quantiles of a uniform distribution were used to determine cluster latent values and unit parameters. The results illustrate that for low intra-class correlation coefficients, the relative increase in MSE can be dramatic. Once again, for low sampling fractions, similar patterns in MSE are evident for (3)–(5).

In Fig. 3, we compare predictors of the sample mean, for (3) and (4) in two settings where cluster sizes differ. Predictor (5) is not applicable in such settings. These results are based on simulation studies (with 5000 trials each) that repeat a two-stage sampling process from a finite population. The MSE is estimated by the average squared difference between the predictor and the latent PSU value in each case. In the left column, cluster sizes differ by 10-fold, with sample sizes for clusters proportional to the cluster size. The results illustrate the performance of the predictors for different sampling fractions. The right column in Fig. 3 compares the MSE of predictors when the sample size per cluster is constant.

5. Discussion

The expanded finite population mixed model uses a larger set of $2N^2$ random variables than the $\mathbb{N}$ random variables typically used in superpopulation models or in the finite population mixed model of Stanek and Singer (2004). These random variables are fewer than the $\mathbb{N}^2$ random variables resulting from an expansion that retains the identity of units and SSUs, and even fewer than the very general representation of the model used by Godambe (1955). We show that this intermediate set of random variables allows a clear representation of a two-stage sample, while accounting for details on different cluster and sample sizes. Other approaches do not appear to connect the potentially observable data to the random variables in the stochastic model. Since more than one finite population mixed model can be used, we have shown how they can be compared by considering them in a hierarchy, and identifying whether the additional set of orthogonal random variables adds to the information about the target quantity. Further reductions in the number of random variables from the expanded finite population mixed model were considered (Appendix B), each of which lead to loss of information.

It is valuable to note that these results depend on selection of the target quantity. For example, if there is interest in the relationship between two variables among units (in a cluster), the collapsed expanded set of $2N^2$ random variables is not likely to be sufficient.

The BLUP obtained under the new model offers substantial gains over previous predictors. These gains are likely mitigated by the need to estimate shrinkage constants for use in practical setups. Simulation studies comparing the performance of the empirical predictors (3)–(5) in the equal cluster size/sample size settings indicate some loss in efficiency, but with a similar ordering of MSE (San Martino et al., 2008). Limited simulation studies have been conducted using the expanded model predictor and have indicated that there is a greater loss in the MSE of (10) relative to the other predictors. Iterative estimation procedures may be possible, and are currently being investigated. This area requires more study.

Appendix A.

Consider the partitioned expanded finite population mixed model in (9) where

$$\mathbf{X}_I = \left[\frac{1}{N}\mathbf{1}_n \otimes \left(\overset{N}{\underset{s=1}{\oplus}} w_s m_s \right) \right] \text{ and } \mathbf{X}_{II} = \left(\begin{array}{c} \dfrac{1}{N}\mathbf{1}_{N-n} \otimes \left(\overset{N}{\underset{s=1}{\oplus}} w_s m_s \right) \\ \hline \dfrac{1}{N}\mathbf{1}_N \otimes \left(\overset{N}{\underset{s=1}{\oplus}} w_s \left(M_s - m_s \right) \right) \end{array} \right),$$

and the random effects are given by

$$
\left[E_{\xi_2}\!\left(\begin{array}{c}\overset{\leftrightarrow}{\mathbf{Y}}_{wI}\\[2pt]\overset{\leftrightarrow}{\mathbf{Y}}_{wII}\end{array}\right) - E_{\xi_1\xi_2}\!\left(\begin{array}{c}\overset{\leftrightarrow}{\mathbf{Y}}_{wI}\\[2pt]\overset{\leftrightarrow}{\mathbf{Y}}_{wII}\end{array}\right)\right] = \left(\begin{array}{c} \mathbf{I}_n \otimes\!\left(\bigoplus\limits_{s=1}^{N} f_s d_s\right)\operatorname{vec}\!\left(\mathbf{U}_I - E_{\xi_1}\left(\mathbf{U}_I\right)\right) \\ \hline \mathbf{I}_{N-n}\otimes\!\left(\bigoplus\limits_{s=1}^{N} f_s d_s\right)\operatorname{vec}\!\left(\mathbf{U}_{II} - E_{\xi_1}\left(\mathbf{U}_{II}\right)\right) \\ \hline \mathbf{I}_N \otimes\!\left(\bigoplus\limits_{s=1}^{N}\left(1-f_s\right)d_s\right)\operatorname{vec}\!\left[\mathbf{U} - E_{\xi_1}\left(\mathbf{U}\right)\right] \end{array}\right),
$$

with

$$
\left(\begin{array}{c}\overset{\leftrightarrow}{\mathbf{E}}_{wI}\\[2pt]\overset{\leftrightarrow}{\mathbf{E}}_{wII}\end{array}\right) = \left(\begin{array}{c}\overset{\leftrightarrow}{\mathbf{Y}}_{wI}\\[2pt]\overset{\leftrightarrow}{\mathbf{Y}}_{wII}\end{array}\right) - E_{\xi_2}\!\left(\begin{array}{c}\overset{\leftrightarrow}{\mathbf{Y}}_{wI}\\[2pt]\overset{\leftrightarrow}{\mathbf{Y}}_{wII}\end{array}\right),
$$

and $\mathbf{U} = (\mathbf{U}_I\ \ \mathbf{U}_{II})$, $\mathbf{U}_I = ((\mathbf{U}_i)) = (\mathbf{U}_1\ \ \mathbf{U}_2\ \ \cdots\ \ \mathbf{U}_n)$ and $\mathbf{U}_{II} = ((\mathbf{U}_i)) = (\mathbf{U}_{n+1}\ \ \mathbf{U}_{n+2}\ \ \cdots\ \ \mathbf{U}_N)$.

The corresponding covariance matrix is given by

$$
\operatorname{var}\!\left(\overset{\leftrightarrow}{\mathbf{Y}}_{wp}\right) = \frac{1}{N-1}\left(\begin{array}{c|c} \mathbf{P}_N \otimes\!\left[\left(\bigoplus\limits_{s=1}^{N} f_s d_s\right)\mathbf{P}_N\!\left(\bigoplus\limits_{s=1}^{N} f_s d_s\right)\right] & \mathbf{P}_N \otimes\!\left[\left(\bigoplus\limits_{s=1}^{N} f_s d_s\right)\mathbf{P}_N\!\left(\bigoplus\limits_{s=1}^{N}\left(1-f_s\right)d_s\right)\right] \\ \hline \mathbf{P}_N \otimes\!\left[\left(\bigoplus\limits_{s=1}^{N}\left(1-f_s\right)d_s\right)\mathbf{P}_N\!\left(\bigoplus\limits_{s=1}^{N} f_s d_s\right)\right] & \mathbf{P}_N \otimes\!\left[\left(\bigoplus\limits_{s=1}^{N}\left(1-f_s\right)d_s\right)\mathbf{P}_N\!\left(\bigoplus\limits_{s=1}^{N}\left(1-f_s\right)d_s\right)\right] \end{array}\right)
$$
$$
+ \left(\begin{array}{c|c} \mathbf{I}_N & -\mathbf{I}_N \\ \hline -\mathbf{I}_N & \mathbf{I}_N \end{array}\right) \otimes\!\left(\bigoplus\limits_{s=1}^{N} v_{se}^2\right)
$$

(see a. http://www.umass.edu/cluster/ for details) which we partition as

$$
\operatorname{var}_{\xi_1\xi_2}\!\left(\begin{array}{c}\overset{\leftrightarrow}{\mathbf{Y}}_{wI}\\[2pt]\overset{\leftrightarrow}{\mathbf{Y}}_{wII}\end{array}\right) = \left(\begin{array}{cc}\mathbf{V}_I & \mathbf{V}_{I,II}\\ \mathbf{V}'_{I,II} & \mathbf{V}_{II}\end{array}\right),\ \ \text{where}\ v_{se}^2 = f_s(1-f_s)\frac{M_s w_s^2 \sigma_s^2}{N}
$$

(see b. http://www.umass.edu/cluster/ for details). Letting $\mathbf{L}$ be a vector of constants, it follows that $\mathbf{L}'\overset{\leftrightarrow}{\mathbf{Y}}_{wI} - T = ((\mathbf{L}' - \mathbf{g}'_I)\ \ -\mathbf{g}'_{II})\left(\begin{array}{c}\overset{\leftrightarrow}{\mathbf{Y}}_{wI}\\[2pt]\overset{\leftrightarrow}{\mathbf{Y}}_{wII}\end{array}\right)$, and the unbiased constraint is given by $(\mathbf{L}' - \mathbf{g}'_I)\mathbf{X}_I - \mathbf{g}'_{II}\mathbf{X}_{II} = 0$. Using Lagrange multipliers, we minimize $\operatorname{var}_{\xi_1\xi_2}(\mathbf{L}'\overset{\leftrightarrow}{\mathbf{Y}}_{wI} - T)$ while accounting for the unbiased constraint results and obtain the familiar solution,

$$
\hat{\mathbf{L}} = \mathbf{g}_I + [\mathbf{V}_I^{-1} - \mathbf{V}_I^{-1}\mathbf{X}_I(\mathbf{X}'_I\mathbf{V}_I^{-1}\mathbf{X}_I)^{-1}\mathbf{X}'_I\mathbf{V}_I^{-1}]\mathbf{V}_{I,II}\mathbf{g}_{II} + \mathbf{V}_I^{-1}\mathbf{X}_I(\mathbf{X}'_I\mathbf{V}_I^{-1}\mathbf{X}_I)^{-1}\mathbf{X}'_{II}\mathbf{g}_{II}.
$$

This result simplifies to

$$
\hat{\mathbf{L}} = \left(\mathbf{P}_n\mathbf{c}_I \otimes \left[\bigoplus\limits_{s=1}^{N}\frac{k_s^*}{f_s}\right]\mathbf{1}_N\right) + \frac{N}{n}\bar{c}\left[\mathbf{1}_n \otimes \left(\bigoplus\limits_{s=1}^{N}\frac{1}{f_s}\right)\mathbf{1}_N\right] + \frac{N-n}{n}\bar{c}_{II}(\mathbf{1}_n \otimes \mathbf{1}_N)
$$

(see c. http://www.umass.edu/cluster/ for details) where $k_s^* = k_s - \frac{k_s}{d_s}\frac{1}{N}\sum_{s^*=1}^{N}((1-k_{s^*})/(1-\bar{k}))d_{s^*}$, $k_s = f_s^2 d_s^2/(f_s^2 d_s^2 + (N-1)v_{se}^2)$ (see d. http://www.umass.edu/cluster/ for details), $\bar{k} = \frac{1}{N}\sum_{s=1}^{N}k_s$, $\bar{c}_{II} = \frac{1}{N-n}\sum_{i=n+1}^{N}c_i$ and $\bar{c} = \frac{1}{N}\sum_{i=1}^{N}c_i$. The predictor $\hat{T}_p = \hat{\mathbf{L}}'\overset{\leftrightarrow}{\mathbf{Y}}_{wI}$ can then be expressed as (10) (see e. http://www.umass.edu/cluster/ for details).

Appendix B.

We discuss whether several other plausible reductions in the dimension of the set of expanded random variables (given by $N\mathbb{N}$), including a reduction to the set of $2N$ random variables used by Stanek and Singer (2004), may be considered without loss of information. First, it is natural to consider whether it is sufficient to predict $T = \mathbf{g}'_p \overset{\leftrightarrow}{\mathbf{Y}}_{wp}$ using the $2N$ collapsed random variables defined by $\mathbf{Y}_w = \mathbf{C}^{*\prime} \overset{\leftrightarrow}{\mathbf{Y}}_{wp}$. This set of random variables is similar to that used by Stanek and Singer (2004) for a population with equal size clusters and equal size samples per cluster with no response error. Since $\mathbf{g}^{*\prime} = \mathbf{g}'_p[\mathbf{C}^*(\mathbf{C}^{*\prime}\mathbf{C}^*)^{-1}] = \mathbf{1}'_2 \otimes \mathbf{c}'$, the target $T = \mathbf{g}^{*\prime} \overset{\leftrightarrow}{\mathbf{Y}}_w$ defines a linear combination of PSU means.

First, let $\hat{T} = \hat{\mathbf{L}}' \mathbf{Y}_{w\mathrm{I}}$ where $\hat{\mathbf{L}}$ represents an $n \times 1$ vector of constants, and $\mathbf{Y}_{w\mathrm{I}}$ represents the first n random variables (corresponding to the sample) in $\mathbf{Y}_w$. In this case, the bias $E_{\xi_1\xi_2}(\hat{\mathbf{L}}'\mathbf{Y}_{w\mathrm{I}} - T) = (\hat{\mathbf{L}}'\mathbf{1}_n)m_s - (\mathbf{c}'\mathbf{1}_N)(M_s - m_s)$ is zero only if sampling of clusters is conducted with probability proportional to size (PPS) (see a. http://www.umass.edu/cluster/).

Now assume a PPS sampling scheme, and notice that since $\mathbf{g}'_p(\mathbf{P}_{\mathbf{C}^*} \overset{\leftrightarrow}{\mathbf{Y}}_{wp}) = 0$, we may write $T = \mathbf{g}^{*\prime}\mathbf{Y}_w + \mathbf{g}'_p(\mathbf{P}_{\mathbf{C}^*} \overset{\leftrightarrow}{\mathbf{Y}}_{wp})$. Letting $\hat{B} = \hat{\mathbf{b}}'(\mathbf{I}_n \otimes \mathbf{P}_N) \overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{I}}$ be a linear unbiased predictor of $\mathbf{g}'_p(\mathbf{P}_{\mathbf{C}^*} \overset{\leftrightarrow}{\mathbf{Y}}_{wp})$ based on the sample part of $\mathbf{P}_{\mathbf{C}^*} \overset{\leftrightarrow}{\mathbf{Y}}_{wp}$ given by $(\mathbf{I}_n \otimes \mathbf{P}_N) \overset{\leftrightarrow}{\mathbf{Y}}_{w\mathrm{I}}$, the predictor $\hat{T}$ will be optimal if and only if $E_{\xi_1\xi_2}[(\hat{T} - T)\hat{B}] = 0$. Simplifying this expectation, we find

$$
\begin{aligned}
E_{\xi_1\xi_2}[(\hat{T} - T)\hat{B}] = {} & [f(f\hat{\mathbf{L}}' - \mathbf{c}'_\mathrm{I}) \otimes \mathbf{1}'_N]\left[\mathbf{I}_n \otimes \left(\frac{1}{N-1}\left(\overset{N}{\underset{s=1}{\oplus}} d_s\right)\mathbf{P}_N\left(\overset{N}{\underset{s=1}{\oplus}} d_s\right)\mathbf{P}_N\right)\right]\hat{\mathbf{b}} \\
& + \left[\frac{f(1-f)}{N}(\hat{\mathbf{L}}' + \mathbf{c}'_\mathrm{I}) \otimes \left[\mathbf{1}'_N\left(\overset{N}{\underset{s=1}{\oplus}} M_s w_s^2 \sigma_s^2\right)\mathbf{P}_N\right]\right]\hat{\mathbf{b}},
\end{aligned}
$$

where $d_s = M_s w_s \mu_s$, $\mathbf{c} = (\mathbf{c}'_\mathrm{I} \; \mathbf{c}'_\mathrm{II})'$, $\mathbf{c}_\mathrm{I}$ is an $n \times 1$ vector and f denotes the common sampling fraction (see b. http://www.umass.edu/cluster/). This expression is not equal to zero, even when the population consists of equal size clusters with homogeneous variances, and equal size samples are taken from sample clusters. By Theorem 1.1 in Rao and Bellhouse (1978), this result implies that some efficiency is lost in prediction when collapsing $\overset{\leftrightarrow}{\mathbf{Y}}_{wp}$ to $\mathbf{Y}_w$. The predictor based on $\overset{\leftrightarrow}{\mathbf{Y}}_{wp}$ will have smaller MSE than the predictor based on $\mathbf{Y}_w$, even in the settings considered by Stanek and Singer (2004) with no response error when clusters are of equal size, and equal size samples are selected.

References

Bolfarine, H., Zacks, S., 1992. Prediction Theory for Finite Populations. Springer-Verlag, New York.

Brown, H., Prescott, R., 1999. Applied Mixed Models in Medicine. Wiley, New York.

Bryk, A.S., Raudenbush, S.W., 2002. Hierarchical Linear Modeling: Applications, Data Analysis Methods. second ed. Sage Publishing, Thousand Oaks, CA.

Cluster web site: ⟨http://www.umass.edu/cluster/⟩. (Details for Section 3.1 in document c06ed54.doc; Details for Section 3.2: a. c07ed15.doc, page 41. b. c07ed15.doc from page 41. c. c07ed25.doc, page 3; Details of Appendix A: a. c07ed27.doc, page 5. b. c07ed15.doc, from page 41. c. c06ed56.doc, from page 42 and, and c07ed01.doc, from page 1. d. c07ed15.doc, page 41. e. c07ed01.doc, page 6. ; Details of Appendix B: a. c07ed29.doc, pages 8. b. c07ed29.doc, page 19).

Cochran, W., 1977. Survey Sampling. Wiley, New York.

Demidenko, E., 2004. Mixed models: Theory and Application. Wiley, New York.

Diggle, P.L., Heagerty, P., Liang, K.Y., Zeger, S., 2002. Analysis of Longitudinal Data. Oxford University Press,

Godambe, V.P., 1955. A unified theory of sampling from finite populations. J. Roy. Statist. Soc. B 17, 269–278.

Goldberger, A.S., 1962. Best linear unbiased prediction in the generalized linear regression model. J. Amer. Statist. Assoc. 57, 369–375.

Graybill, F.A., 1983. Matrices with Applications in Statistics. Wadsworth International, Belmont, CA.

Harville, D.A., 1978. Alternative formulations and procedures for the two-way mixed model. Biometrics 34, 441–453.

Hedayat, A.S., Sinha, B.K., 1991. Design and Inference in Finite Population Sampling. Wiley, New York.

Henderson, C.R., 1984. Applications of Linear Models in Animal Breeding. University of Guelph, Guelph, Canada.

Jiang, J., 2007. Linear and Generalized Linear Mixed Models and their Applications. Springer, New York.

Laird, N.M., Ware, J.H., 1982. Random-effects models for longitudinal data. Biometrics 38, 963–974.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., Schabenberger, O., 2006. SAS for Mixed Models. Second ed. SAS Institute, Cary, NC.

McCulloch, C.E., Searle, S.R., 2001. Generalized, Linear, and Mixed Models. Wiley, New York.

McLean, R.A., Sanders, W.L., Stroup, W.W., 1991. A unified approach to mixed linear models. The American Statistician 45 (1), 54–64.

Rao, J.N.K., Bellhouse, D.R., 1978. Optimal estimation of a finite population mean under generalized random permutation models. J. Statist. Plann. Inference 2, 125–141.

Robinson, G.K., 1991. That BLUP is a good thing: the estimation of random effects. Statist. Sci. 6 (1), 15–51.

Royall, R.M., 1976. The linear least-squares prediction approach to two-stage sampling. J. Amer. Statist. Assoc. 71, 657–664.

San Martino, S., Singer, J.M., Stanek, E.J. III, 2008. Performance of balanced two-stage empirical predictors of realized cluster latent values from finite populations: a simulation study. Comput. Statist. Data Anal. 52, 2199–2217.

Scott, A., Smith, T.M.F., 1969. Estimation in multi-stage surveys. J. Amer. Statist. Assoc. 64 (327), 830–840.

Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. Wiley, New York.

Singer, J.D., Willett, J.B., 2003. Applied Longitudinal Data Analysis. Oxford Press, New York.

Stanek III, E.J., Singer, J.M., 2004. Predicting random effects from finite population clustered samples with response error. J. Amer. Statist. Assoc. 99, 119–130.

Stanek III, E.J., Singer, J.M., Lencina, V.B., 2004. A unified approach to estimation and prediction under simple random sampling. J. Statist. Plann. Inference 121, 325–338.

Valliant, R., Dorfman, H.A., Royall, R.M., 2000. Finite Population Sampling and Inference. Wiley, New York.

Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.