Equation Chapter 1 Section 1 Predicting Random Effects from a Finite Population of Unequal Size Clusters based on Two-Stage Sampling

Edward J. Stanek III
Department of Public Health
401 Arnold House
University of Massachusetts
715 North Pleasant Street
Amherst, MA 01003-9304 USA
stanek@schoolph.umass.edu

Julio M. Singer
Departamento de Estatística
Universidade de São Paulo
São Paulo, Brazil
jmsinger@ime.usp.br
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. Best linear unbiased predictors developed from mixed models are widely used, but their development requires distributional assumptions or an infinite population framework. When the number and size of clusters is finite, super population models have been used to predict the contribution of the unobserved units to a realized cluster mean. Recently, predictors developed from a two-stage sampling model have been shown to out-perform the model-based predictors. However, the random permutation model underlying these predictors is limited to settings where cluster sizes and sample sizes per cluster are equal.

We present a new two-stage sampling model that can be used when cluster sizes and sample sizes per cluster differ. The model expands the set of random variables from the two-stage sampling model to accommodate the different size clusters. The resulting best linear unbiased predictor out performs any of the competing predictors, even when clusters and sample sizes per cluster are equal. The reduction in expected mean squared error relative to predictors developed under mixed model or super population model assumptions is likely due to the specificity of the expanded random permutation model to the problem. This suggests faithfully capturing the stochastic aspects of a problem is more important than simplifying assumptions in developing optimal solutions. Many other problems may be amenable to improved solutions based on extension of this approach.
Contact Address:
Edward J. Stanek III
Department of Public Health
401 Arnold House
University of Massachusetts at Amherst
Amherst, Ma. 01002

Phone: 413-545-3812
Fax: 413-545-1645
Email: stanek@schoolph.umass.edu

KEYWORDS: superpopulation, best linear unbiased predictor, random permutation, optimal estimation, design-based inference, mixed models.

ACKNOWLEDGEMENT. This work was developed with the support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil and the National Institutes of Health (NIH-PHS-R01-HD36848, R01-HL071828-02), USA.
1. INTRODUCTION

Prediction of realized random effects is an important problem that has broad application. In its simplest form, the problem corresponds to predicting the mean of a realized cluster in a two stage simple random sample. There are many applications where such prediction is of interest. Henderson (1984) was interested in predicting a cow’s milk production, and studied lactation records of dairy cows. In the current U.S. educational environment, high school standardized test scores are of interest, with schools having high stakes in the results due to their use in resource allocation. The strength of a family’s emotional support is of interest in studies of coping with subsequent illness in aging (Yorgason, et al. 2006). A patient’s biomarker level may be important in quantifying risk of disease or subsequent intervention.

Although applications vary widely, each problem can be defined using a common clustered population framework, where the target parameter can be defined as the average expected response of units in a realized cluster. In the examples, a cluster may correspond to a cow, a school, a family, or a patient, while the units may correspond to lactation periods, students, family members, or time periods, respectively. Ideally, the number and size of clusters is known, since the practical relevance and interpretation of the statistical inference is enhanced by a clear problem definition.

It is rarely possible to observe the expected response on all units in each cluster. Instead, response is observed on a subset of units for a subset of clusters. An important question is how to use such information to guess the expected response of a specified
cluster. Without further structure and assumptions, the universe of such guesses is limitless, and determining a best guess seems impossible. We bound this problem by assuming a link between the observed data and the population via a probability model. The probability model is a two-stage sampling model that represents the population as a set of random variables, and considers the observed response to be a realization of a subset of these random variables. We refer to this model as a random permutation (RP) model. The model will have face validity when the observed response is the result of a two stage cluster sample. In other settings, the model provides a simple starting point for statistical inference.

The choice of a probability model will impact the problem definition, and interpretation of solutions. First, although clusters are identifiable in the finite population, this identifiability is lost in the typical representation of random variables in a probability model. In a probability model, the subscript that identifies a cluster in the population is usually replaced by a subscript representing the position of a selected cluster in the sample (Cassel, Sarndal, and Wretman 1979). A similar substitution occurs for the subscript identifying units. These differences imply that the expected response for the first cluster in a population listing is not the same as the expected response for the first sample cluster, which we refer to as the first primary sampling unit, or PSU. In a similar way, we refer to second stage units in the probability model as secondary sampling units, or SSUs. The cluster that is assigned to a PSU in a realized sample is the realized cluster, and the difference between the expected response for that cluster and the population
average is the realized random effect. These difference between a cluster and a PSU is often a source of confusion in interpreting predictors of realized random effects.

Probability models usually include the notion of a PSU, and the expected value of a realized PSU. A problem occurs, however, when trying to represent SSUs in a PSU when cluster sizes differ. This problem is important since the mean of a realized PSU can be expressed as a weighted sum of two parts: the average expected response of sample SSUs in the PSU, and the average expected response of the remaining non-sampled SSUs. If the expected response for sample SSUs is known, predicting the expected response of the remaining SSUs is the objective of the statistical inference. When clusters are not the same size, it is not possible to identify the remaining non-sampled SSUs with the usual probability models. To sidestep this difficulty, model based approaches assume the number of SSUs is infinitely large (as in the mixed model approaches of Goldberger 1962; Henderson 1984, McLean, Sanders and Stroup 1991; Robinson 1991; and McCulloch and Searle 2001), or condition on cluster size (as in the superpopulation model based approaches of Scott and Smith 1969; Ghosh and Lahiri 1987; Bolfarine and Zachs 1992; Valliant, Dorfman and Royall, 2000). In the first case, to avoid the problem of identifying the remaining SSUs, a solution is developed for a different problem. In the second case, the conditioning on a realized cluster’s size skips over the awkward problem that the cluster size is linked to identifiable clusters, and clusters are not identifiable in the probability model.

Model based approaches to predicting a realized random effect are characterized by a gap between the identifiable clusters and units in the finite population and the
probability model from which the solution is posed. Since multiple probability models can be posited, different ‘optimal’ solutions can be claimed for the same basic problem. Each solution suffers from the limitation that the underlying model can not be connected back to the identifiable finite population. Thus, while each solution is optimal for its theoretical framework, none of the frameworks match reality, and it unclear which optimal solution is best.

We present a new expanded two stage random permutation model that overcomes some of the limitations of previous probability models. A key feature of the expanded model is simultaneously retaining the cluster identity and the PSU position in the model, while distinguishing for a PSU the relevant contribution of sample SSUs, and non-sampled SSUs to a target random variable. The model extends the expanded model for simple random sampling of Stanek, Singer, and Lencina (2004) to the two-stage RP model used by Stanek and Singer (2004), while allowing for possible unequal size clusters.

Beginning with an explicit representation of random variables underlying a two stage random permutation of a finite population, we show that when predicting a PSU mean via a linear unbiased predictor, the expanded model can not be further reduced without loss of information. Since the expanded model retains this information, the best linear unbiased predictor (developed in an identical manner as the development in Stanek and Singer (2004)), has smaller expected MSE than any previously developed predictors (including commonly used mixed model predictors, super population model predictors). We characterize the extent of the reduction in the theoretical expected mean squared error,
and illustrate results for empirical predictors via simulations relative to mixed model and super population model predictors. We conclude by highlighting model features that have consequence in extending this work and related work that offer promising possibilities for future improvement.

2. AN EXPANDED RP MODEL FOR A FINITE CLUSTERED POPULATION

We first define notation and terminology for a clustered finite population. Let a finite population be defined by a listing of \( M_s \) units labeled \( t = 1, \ldots, M_s \) in each of \( N \) clusters, labeled \( s = 1, \ldots, N \), where the non-stochastic response for unit \( t \) in cluster \( s \) is given by \( y_{st} \). We assume that the response for a unit can be observed without error, and corresponds to the expected value for the unit. The finite population parameters correspond to the mean and variance of cluster \( s \), \( s = 1, \ldots, N \), are defined by

\[
\mu_s = \frac{1}{M_s} \sum_{t=1}^{M_s} y_{st} \quad \text{and} \quad \left( \frac{M_s - 1}{M_s} \right) \sigma^2_s = \frac{1}{M_s} \sum_{t=1}^{M_s} (y_{st} - \mu_s)^2
\]

(where we use the survey sampling definition of the parameter \( \sigma^2_s \)). Similarly, the population mean, and the variance between cluster means are defined as

\[
\mu = \frac{1}{N} \sum_{s=1}^{N} \mu_s \quad \text{and} \quad \left( \frac{N - 1}{N} \right) \sigma^2 = \frac{1}{N} \sum_{s=1}^{N} (\mu_s - \mu)^2.
\]

Using these parameters, we represent the potentially observable response for unit \( t \) in cluster \( s \) as

\[
y_{st} = \mu + \beta_s + e_{st} \quad (1)
\]
where $\beta_s = (\mu_s - \mu)$ is the deviation of the mean for cluster $s$ from the overall mean, and $\epsilon_{st} = (y_{st} - \mu_s)$ is the deviation of unit $t$'s response (in cluster $s$) from the mean for cluster $s$. Defining $y = \left( y'_1 \quad y'_2 \quad \cdots \quad y'_N \right)'$ where $y_s = \left( y_{s1} \quad y_{s2} \quad \cdots \quad y_{sM_s} \right)'$, the model can be summarized as

$$y = X\mu + Z\beta + \epsilon$$

where $X = 1_N$, $Z = \bigoplus_{s=1}^{N} I_{M_s}$, $N = \sum_{s=1}^{N} M_s$, and $\beta = (\beta_1 \quad \beta_2 \quad \cdots \quad \beta_N)'$. Here, $1_n$ is an $a \times 1$ column vector of ones, $\bigoplus_{s=1}^{N} A_s$ denotes a block diagonal matrix with blocks given by $A_s$ (Graybill 1983), and $\epsilon$ is defined similarly to $y$. None of the terms in the model are random variables.

### 2.1. Random Variables and The Two Stage Random Permutation Model

We explicitly define a vector of random variables that represents a two stage random permutation of the population. Assuming that each realization of the permutation is equally likely, with probability $\frac{1}{N!\prod_{s=1}^{N} M_s!}$, the random variables formally represent two-stage sampling (Cochran 1977). We assume that the sample clusters are in the first $n$ positions in a permutation of clusters. Similarly, we assume that the sample units in
cluster \( s \) correspond to the units in the first \( m_s \) positions in a permutation of the cluster’s units.

When all clusters are of equal size such that \( M_s = M \) for all \( s = 1, \ldots, N \), Stanek and Singer (2004) define indicator random variables to explicitly represent a random vector of dimension \( NM \times 1 \) corresponding to a two-stage permutation of the population. We follow a similar strategy when clusters sizes differ, but note two differences. First, in order to clearly identify the PSUs and SSUs in the vector representing a random permutation, we expand the number of random variables from \( N \) to \( 2N \). Second, we include a set of fixed known weights associated with the SSUs in a cluster. Different choices of weights allow us to easily change the definition of a target parameter (such as a mean, total, or a cluster regression parameter). We define the weighted expanded random variables, and illustrate the need for the expansion next.

Sample indicator random variables relate \( y_{st} \) (response for unit \( t \) in cluster \( s \)) to the response for a unit in position \( j \) of a cluster in position \( i \) in a two stage permutation of clusters and units. We define \( U_{jt}^{(s)} \) as an indicator random variable that takes on a value of one when SSU \( j \) in cluster \( s \) is unit \( t \), and zero otherwise, and use it to represent the random variable corresponding to SSU \( j \) in cluster \( s \) given by \( \bar{Y}_{sj} = \sum_{t=1}^{M_s} U_{jt}^{(s)} y_{st} \). Let \( w_{sj} \) denote a fixed non-stochastic weight for \( s = 1, \ldots, N \), \( j = 1, \ldots, M_s \), and define the weighted response as \( \bar{y}_{wj} = w_{sj} \bar{y}_{sj} \). For example, PSU totals are defined by setting \( w_{sj} = 1 \).
for all $s = 1, \ldots, N$, $j = 1, \ldots, M_s$; PSU means are defined by setting $w_{sj} = \frac{1}{M_s}$ for all $s = 1, \ldots, N$, $j = 1, \ldots, M_s$.

We represent response for a permutation of units in cluster $s$ by the $M_s \times 1$ vector

$$
\tilde{y}_{ws} = \left( \bigoplus_{j=1}^{M_s} w_{sj} \right) U^{(s)} y_s = \left( \left( \tilde{y}_{wjs} \right) \right), \text{ where } U^{(s)} = \begin{pmatrix} U_1^{(s)} & U_2^{(s)} & \cdots & U_{M_s}^{(s)} \end{pmatrix} \text{ is an } M_s \times M_s
$$

matrix of indicator random variables with columns $U^{(s)} = \begin{pmatrix} U_1^{(s)} & U_2^{(s)} & \cdots & U_{M_s}^{(s)} \end{pmatrix}$.

A permutation of clusters is defined using the indicator random variables $U_{is}$ for $i = 1, \ldots, N$ and $s = 1, \ldots, N$, that take on a value of one when PSU $i$ is cluster $s$, and a value of zero otherwise. We summarize these random variables in the $N \times N$ matrix $U = \left( \left( U_{is} \right) \right)$. If all clusters were equal in size, we could represent the random variable corresponding to PSU $i$ and SSU $j$ by $\sum_{s=1}^{N} U_{ij} \tilde{y}_{wjs}$. Note that when cluster sizes differ, this expression is defined only when $j \leq M_s$, since $j$ indexes the SSUs in cluster $s$. Mixed model and superpopulation model representations of two stage cluster sampling do not formally account for this range restriction in linking the random variables to the finite population.

We directly account for different size clusters, and avoid the requirement of range constraints for subscripts by expanding the number of random variables used to represent a permutation of clusters and units. For PSU $i$, the expanded random variables are defined by the $N \times 1$ vector $\hat{Y}_{wi} = \left( \left( U_{i1} \tilde{y}_{w1} \right) U_{i2} \tilde{y}_{w2} \cdots U_{iN} \tilde{y}_{wN} \right)^\prime$, with a
two stage random permutation of the population represented by the \(\mathbf{N} \times 1\) vector,

\[
\mathbf{\tilde{Y}}_w = \left( (\mathbf{\tilde{Y}}_{w_1}) \right) = \left( \mathbf{\tilde{Y}}_{w_1}' \quad \mathbf{\tilde{Y}}_{w_2}' \quad \ldots \quad \mathbf{\tilde{Y}}_{w_N}' \right)' .
\]

An element of \(\mathbf{\tilde{Y}}_w\) is given by

\[
U_a w_{sj} \mathbf{\tilde{Y}}_{sj} = U_{ia} w_{sj} \sum_{r=1}^{M_i} U_j^{(s)} y_{sr} .
\]

A simple example illustrates the notation. Suppose the population consists of three clusters \((N = 3)\), where the first two clusters have two units \((M_1 = M_2 = 2)\), the third cluster has three units \((M_3 = 3)\), \(\mathbf{N} = 7\), and \(w_{sj} = 1\) for all \(s = 1, \ldots, N\), \(j = 1, \ldots, M_s\).

We represent a permutation of units for cluster \(s\) by \(\mathbf{\tilde{Y}}_s = \mathbf{\tilde{Y}}_{ws} , s = 1, \ldots, 3\). Suppose the first permutation of clusters results in clusters \(s = 1, s = 2, \text{ and } s = 3\) in positions \(i = 1, \ldots, 3\) respectively, while a second permutation results in clusters \(s = 3, s = 2, \text{ and } s = 1\) in positions \(i = 1, \ldots, 3\) respectively. The representation of the random variables realized by the first permutation of PSUs is the random vector \(\left( \mathbf{\tilde{Y}}_1' \quad \mathbf{\tilde{Y}}_2' \quad \mathbf{\tilde{Y}}_3' \right)'\), and by the second permutation of PSUs is the random vector \(\left( \mathbf{\tilde{Y}}_3' \quad \mathbf{\tilde{Y}}_2' \quad \mathbf{\tilde{Y}}_1' \right)'\). Although both vectors are of dimension \(\mathbf{N} \times 1\), the third SSU in the first permutation is in PSU \(i = 1\), while the third SSU in the second permutation is in PSU \(i = 2\). The position of a SSU in the permuted population is not sufficient to retain the identity of the PSU for the SSU. In contrast, using the expanded random variable representation, the random variables realized by the first permutation of PSUs are represented by

\[
\left( \mathbf{\tilde{Y}}_1' \quad \mathbf{0}_2' \quad \mathbf{0}_3' \quad \mathbf{0}_2' \quad \mathbf{\tilde{Y}}_2' \quad \mathbf{0}_3' \quad \mathbf{0}_2' \quad \mathbf{\tilde{Y}}_3' \right)' ,
\]

while those realized by the second
permutation are represented by \( \left( 0_2', 0_2', 0_2', 0_3', 0_3', 0_3', \hat{Y}_2', \hat{Y}_2', \hat{Y}_3', \hat{Y}_3', \hat{Y}_3', \hat{Y}_3' \right)' \), where \( 0_a \) is an \( a \times 1 \) vector with all elements equal to zero. This notation preserves the identity of the PSU for each SSU, avoiding the ambiguity that arises in mixed models and superpopulation models.

2.2 A Mixed Effect Model for the Expanded Random Variables

We represent a mixed model for the expanded RP model representing expectation with respect to permutations of clusters with the subscript \( \xi_1 \) and expectation with respect to permutations of units in a cluster with the subscript \( \xi_2 \). For PSU \( i \), we express:

\[
\tilde{Y}_{wi} = E_{\xi_1} (\bar{Y}_{wi}) + \left[ E_{\xi_1|\xi_0} (\bar{Y}_{wi}) - E_{\xi_1} (\bar{Y}_{wi}) \right] + \tilde{E}_{wi}
\]

where \( E_{\xi_1|\xi_0} (\bar{Y}_{wi}) = \frac{1}{N} \begin{pmatrix} \mathbf{w}_1 \mu_1 \\ \mathbf{w}_2 \mu_2 \\ \vdots \\ \mathbf{w}_N \mu_N \end{pmatrix} \), \( E_{\xi_1} (\bar{Y}_{wi}) = \left( \bigoplus_{s=1}^{N} \mathbf{w}_s \mu_s \right) \mathbf{U}_i \), \( \tilde{E}_{wi} = \bar{Y}_{wi} - E_{\xi_1|\xi_0} (\bar{Y}_{wi}) \), and

\( \mathbf{w}_s = \left( \begin{pmatrix} w_{s1} \\ w_{s2} \\ \vdots \\ w_{sM_s} \end{pmatrix} \right) \). The fixed effects, \( E_{\xi_1} (\bar{Y}_{wi}) \), correspond to a vector of weighted cluster means. The random effects, \( E_{\xi_1|\xi_0} (\bar{Y}_{wi}) - E_{\xi_1} (\bar{Y}_{wi}) \), are defined as the deviation from the fixed effects of the expected response conditional on a realized PSU. The deviation of response from the expected response within a PSU is represented by \( \tilde{E}_{wi} \). We combine these expressions arriving at the expanded RP mixed model given by
\[
\hat{Y}_w = \left[ I_N \otimes \left( \sum_{s=1}^{N} w_s \mu_s \right) \right] \frac{1}{N} + \left[ I_N \otimes \left( \sum_{s=1}^{N} w_s \mu_N \right) \right] \left[ \text{vec}(U) - E_{\xi_i} \left( \text{vec}(U) \right) \right] + \tilde{E}_w . \tag{2}
\]

For this model, \( E_{\xi_i | \tilde{Y}_w} \left( \hat{Y}_w \right) = \left[ I_N \otimes \left( \sum_{s=1}^{N} w_s \mu_s \right) \right] \text{vec}(U) \),

\[
E_{\xi_{\tilde{w}}} \left( \hat{Y}_w \right) = \left[ \frac{1}{N} I_N \otimes I_N \right] \left( \begin{array}{ccc}
w_1 \mu_1 \\
\vdots \\
\vdots \\
w_N \mu_N 
\end{array} \right),
\]

the variance of the random effects is

\[
\text{var}_{\xi_{\tilde{w}}} \left[ \left[ I_N \otimes \left( \sum_{s=1}^{N} w_s \mu_s \right) \right] \left[ \text{vec}(U) - E_{\xi_i} \left( \text{vec}(U) \right) \right] \right] = \frac{1}{N-1} P_N \otimes \left[ \left( \sum_{s=1}^{N} w_s \mu_s \right) P_N \left( \sum_{s=1}^{N} w_s \mu_s \right) \right]
\]

and \( \text{var}_{\tilde{w}} \left( \tilde{E}_w \right) = I_N \otimes \left( \sum_{s=1}^{N} \frac{\sigma^2}{N} M \left( \sum_{j=1}^{M} w_{sj} P_{M_s} \left( \sum_{j=1}^{M} w_{sj} \right) \right) \right) \), where \( P_a = I_a - \frac{1}{a} J_a \) and \( J_a \)
denotes an \( a \times a \) matrix with all elements equal to one.

### 2.3. Defining Random Variables of Interest

Model (2) is an expanded version of a mixed model that retains the identity of clusters, while accounting for a two stage random permutation. Our interest is in predicting various random effects, such as the mean (or total) of a realized cluster. The basic target random variables can be defined by \( T = g_i' \hat{Y}_w \) or \( T_A = g_i' E_{\xi_i} \left( \hat{Y}_w \right) \) where

\( g_i' = e_i' \otimes I_n' \) and \( e_i \) is an \( N \times 1 \) vector with all elements equal to zero, except for element \( i \) which has the value of one. Of principal interest is the setting where \( i \leq n \), such that the cluster of interest is realized in the sample.

Different weights result in different target random variables. We partition

\[
w_s = \left( w_{sl}' \quad w_{sl}' \right)' \]

into a sample and remaining vector for \( s = 1, \ldots, N \), and limit discussion
to settings where \( w_{sl} = w_{sl} 1_{m_s} \) and \( w_{ssl} = w_{ssl} 1_{M_s-m_s} \). When \( w_{sl} = w_{ssl} = \frac{1}{M_s} \), the target parameter corresponds to \( T = T_A = \sum_{s=1}^{N} U_{is}\mu_s \), the mean of PSU \( i \). When \( w_{sl} = w_{ssl} = 1 \), \( T = T_A = \sum_{s=1}^{N} U_{is}M_s\mu_s \) which corresponds to the total of PSU \( i \). As a third illustration, suppose the weights are chosen such that \( w_{sl} = \frac{1}{2m_s} \) and \( w_{ssl} = \frac{1}{2(M_s-m_s)} \). In this setting, \( T_A = \sum_{s=1}^{N} U_{is}\mu_s \), the mean of PSU \( i \), but \( T = \sum_{s=1}^{N} U_{is}\mu_s + \sum_{s=1}^{N} U_{is} \left( \frac{1}{2}(\bar{E}_{sl} + \bar{E}_{ssl}) \right) \) where \( \frac{1}{m_s} \sum_{j=1}^{m_s} \bar{E}_{sj} \) and \( \frac{1}{M_s-m_s} \sum_{j=m_s+1}^{M_s} \bar{E}_{sj} \) and \( \bar{E}_{sj} = U_j^{(s)}\epsilon_s \). Thus, different choice in weights can result in some subtle differences in the target random variables. We focus on predicting the mean of a sample PSU given by \( T = \sum_{s=1}^{N} U_{is}\mu_s \) subsequently, assuming that the weights are given by \( w_{sl} = w_{ssl} = \frac{1}{M_s} \) for \( s = 1, \ldots, N \).

3. PREDICTING A PSU MEAN USING AN EXPANDDED RP MODEL

The basic strategy for developing a predictor under a model-based approach is given in many places (Scott and Smith 1969; Royall 1976; Bolfarine and Zacks 1992; Valliant et al. 2000), and applied in a design-based framework to a balanced two stage
cluster sampling by Stanek and Singer (2004). Stanek and Singer showed that the predictor of a realized PSU mean from the RP model outperforms mixed model and superpopulation model predictors. We develop similar results using the expanded RP model, and illustrate that the resulting predictor of a realized PSU mean has even lower expected MSE relative to the predictor given by Stanek and Singer (2004).

We assume that the elements in the sample portion of \( \bar{Y}_m \) will be observed, and express \( T \) as the sum of two parts, one which is a function of the sample, and the other which is 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, coefficients are evaluated that minimize, the expected value of the MSE given by \( \text{var}_{\xi_i \xi_2} (\hat{T} - T) \). While in theory, an optimal predictor can be expressed following this prediction recipe, in practice, the high dimensionality of the vectors from the expansion of random variables will result in singularities that prevent unique solution. For this reason, we examine linear functions that project the random variables into a lower dimensional space, while still retaining information necessary for an optimal solution of the problem. Notice that we avoid the traditional representation which collapses the expanded random variables to the terms \( Y_p = \sum_{s=1}^{K} \sum_{j=1}^{M_j} U_s U^{(s)} y^t_m \), leading to random dimensions for the vectors of PSU random variables.

3.1. Collapsing Expanded RP Random Variables without Loss of Information
We collapse the expanded random variables without loss of information to simplify the problem making use of Theorem 1.1 of Rao and Bellhouse (1978). The theorem provides a way of determining if a set of random variables are sufficient for a given problem. It states that if we can express the target random variables as the sum of two orthogonal terms, one of which has expectation zero, then the predictor of the non-zero term will be optimal if the expectation of the predictor minus its expected value times the predictor of zero is zero. We first illustrate application of the theorem to simple random sampling, showing that information is lost if the usual random variables are used to predict linear combinations of \( \mathbf{Y} = \left( \left( Y_i \right) \right) \) where \( Y_i = \sum_{s=1}^{N} U_{is} Y_s \). Next, we show how certain collapsing can be made in a two-stage problem without loss of information.

In a simple random sample, we can use this theorem to show that the usual random variables, \( \mathbf{Y} = \left( \left( Y_i \right) \right) \) with \( Y_i = \sum_{s=1}^{N} U_{is} Y_s \), are sufficient for predicting a linear combination of \( \mathbf{Y} \) given by \( P = \mathbf{c}' \mathbf{Y} = \mathbf{L} \bar{\mathbf{Y}} \) where \( \mathbf{L} = \mathbf{c}' \otimes \mathbf{I}_N \), \( \mathbf{c} = \left( \left( c_i \right) \right) \), with \( c_i \) constant, \( \bar{\mathbf{Y}} = \left( \left( \bar{Y}_i \right) \right) \) where \( \bar{Y}_i = (Y_{i1} \ Y_{i2} \ \cdots \ Y_{in})' \) and \( Y_{is} = U_{is} Y_s \). To see this, notice that we can express \( \bar{\mathbf{Y}} = \mathbf{X} \bar{\mathbf{Y}} + \mathbf{P}_x \bar{\mathbf{Y}} \) as the sum of two orthogonal terms defined by

\[
\mathbf{X} = \mathbf{I}_N \otimes \frac{1}{N} \mathbf{J}_N, \quad \mathbf{P}_x = \mathbf{I}_N \otimes \mathbf{P}_N \text{ and } \mathbf{X} \bar{\mathbf{Y}} = \left( \mathbf{I}_N \otimes \frac{1}{N} \mathbf{I}_N \right) \mathbf{Y}.
\]

With these definitions, the target random variable can be expressed as the sum of two orthogonal terms,

\[ P = \mathbf{c}' \mathbf{Y} + \mathbf{L} \mathbf{P}_x \bar{\mathbf{Y}}, \]

where the second term is zero. We define the sample to be the first \( n \) elements of \( \mathbf{Y} \) which we represent by \( \mathbf{Y}_i \), or equivalently as the first \( nN \) elements of \( \bar{\mathbf{Y}} \).
given by $\hat{Y}_f$. Let $C$ denote the class of linear unbiased predictors of $P$ based on $Y_i$, such that the optimal predictor in this class is given by $\hat{T} = \hat{a}Y_i$. Let $C_0$ denote the class of linear unbiased predictors of $B = LP_X\tilde{Y} = 0$ based on $(I_n \otimes P_N)\tilde{Y}_f$ which we express as $\tilde{B}^c = b'(I_n \otimes P_N)\tilde{Y}_f$. Rao and Bellhouse’s theorem states that $\hat{T}$ is optimal for $P = L\tilde{Y}$ if and only if $E\left[\left(\hat{T}_f^c P\right)B\right] = 0$ for all $\tilde{B}$. Since

$$E\left[\left(\hat{T}_f^c T\right)B\right] = \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i (y_i - \mu) \left[\sum_{i=1}^{n} (\hat{a}_{i} - c_i)\tilde{b}_i\right]\right) - \sigma^2 \sum_{i=1}^{n} (a_i - c_i)\tilde{b}_i$$

where $\tilde{b}_i = \frac{1}{N} \sum_{i=1}^{N} \tilde{b}_i$.

We consider predictors that are a linear function of the sample. Consider now an unbiased predictor of $c'Y$ based on and an unbiased predictor of $LP_X\tilde{Y}$ based on a simple random sample of size $n$. We use Rao and Bellhouse’s theorem in this context to show that a linear unbiased predictor based on a simple random sample can not be developed based solely on $Y$.

TO HERE 10/30/2006

In a two-stage RP model, it is not possible to collapse $\tilde{Y}_w$ to the set of random variables usually defined for a mixed model, or superpopulation model without loss of
information when predicting linear combinations of PSU means. We can define, however, a collapsed set of random variables from which optimal predictors can be developed. The collapsed random variables correspond to sample and remainder SSU totals in each PSU/cluster combination. We define the collapsed random variables as

\[
\tilde{\mathbf{C}}' \tilde{\mathbf{Y}}_w \text{ where } \tilde{\mathbf{C}} = \bigoplus_{i=1}^{N} \bigoplus_{s=1}^{N} \tilde{\mathbf{c}}_{si}, \quad \tilde{\mathbf{c}}_{si} = \left( \mathbf{1}_{m_i} \quad \mathbf{0}_{M_s - m_i} \right)' \quad \text{and} \quad \tilde{\mathbf{c}}_{si} = \left( \mathbf{0}_{m_s} \quad \mathbf{1}_{M_s - m_s} \right)'.
\]

Dividing these random variables into a sample and remainder portion, we represent

\[
\tilde{\mathbf{C}}' \tilde{\mathbf{Y}}_w \text{ by } \begin{pmatrix} \tilde{\mathbf{Y}}_{wI} \\ \tilde{\mathbf{Y}}_{wII} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{C}}_I \tilde{\mathbf{Y}}_w \\ \tilde{\mathbf{C}}_II \tilde{\mathbf{Y}}_w \end{pmatrix}, \text{ where } \tilde{\mathbf{C}}_I = \bigoplus_{i=1}^{N} \bigoplus_{s=1}^{N} \mathbf{0}_{m_i} \quad \bigoplus_{i=1}^{N} \bigoplus_{s=1}^{N} \mathbf{0}_{m_s} \quad \bigoplus_{i=1}^{N} \bigoplus_{s=1}^{N} \mathbf{0}_{m_i} \quad \bigoplus_{i=1}^{N} \bigoplus_{s=1}^{N} \mathbf{0}_{m_s} \quad \bigoplus_{i=1}^{N} \bigoplus_{s=1}^{N} \mathbf{0}_{M_s - m_i} \quad \bigoplus_{i=1}^{N} \bigoplus_{s=1}^{N} \mathbf{0}_{M_s - m_s} \quad \bigoplus_{i=1}^{N} \bigoplus_{s=1}^{N} \mathbf{0}_{M_s - m_s} \quad \bigoplus_{i=1}^{N} \bigoplus_{s=1}^{N} \mathbf{0}_{M_s - m_s} \quad \bigoplus_{i=1}^{N} \bigoplus_{s=1}^{N} \mathbf{0}_{M_s - m_s}
\]

Notice that \( \tilde{\mathbf{Y}}_w = \tilde{\mathbf{C}}(\tilde{\mathbf{C}}' \tilde{\mathbf{C}})^{-1} \tilde{\mathbf{C}}' \tilde{\mathbf{Y}}_w + \mathbf{P}_c \tilde{\mathbf{Y}}_w \).

\[
\mathbf{P}_c = \mathbf{I}_{N^2 \times N^2} - \tilde{\mathbf{C}}(\tilde{\mathbf{C}}' \tilde{\mathbf{C}})^{-1} \tilde{\mathbf{C}}' = \bigoplus_{i=1}^{N} \bigoplus_{s=1}^{N} \begin{pmatrix} \frac{1}{m_s} \mathbf{P}_{m_s} \\ \mathbf{0} \end{pmatrix} \quad \frac{1}{M_s - m_s} \mathbf{P}_{M_s - m_s}
\]

orthogonal. Since \( E_{s_i \mid I} (\tilde{\mathbf{Y}}_w) = \left[ \mathbf{1}_N \bigoplus_{i=1}^{N} \mathbf{w}_i \right] \mathbf{vec} (\mathbf{U}) \) and weights defined are of the form \( \mathbf{w}_s = (\mathbf{w}'_s \quad \mathbf{w}''_s)' \), where \( \mathbf{w}_s = w_{si} \mathbf{1}_{m_i}, \mathbf{w}_s = w_{si} \mathbf{1}_{M_s - m_i}, E_{s_i \mid I} \left( \mathbf{P}_c \tilde{\mathbf{Y}}_w \right) = \mathbf{0} \),

This implies that since \( \tilde{\mathbf{Y}}_w = \tilde{\mathbf{C}}(\tilde{\mathbf{C}}' \tilde{\mathbf{C}})^{-1} \tilde{\mathbf{C}}' \tilde{\mathbf{Y}}_w + \mathbf{P}_c \tilde{\mathbf{Y}}_w \), we can develop an optimal
estimator or predictor solely based on $\begin{pmatrix} \hat{C} \hat{C}' \hat{C} \end{pmatrix} ^{-1} \hat{C}' \hat{Y}$. We express this as

$$\begin{pmatrix} \hat{C} \hat{C}' \hat{C} \end{pmatrix} ^{-1} \begin{pmatrix} \hat{Y} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_{i} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} \sum_{k=1}^{N} \tilde{c}_{ik} \end{pmatrix} + \begin{pmatrix} \sum_{i=1}^{N} m_{s} \end{pmatrix}.$$

of SSUs se random variables It is possible, however, to

, who identify conditions under which such collapsing, Theorem 1.1) describe criteria
under which an optimal uniform minimum variance estimator can be obtained based on a projection of random variables. We review the basic idea using an expanded model for a simple random sample discussed by Stanek, Singer, and Lencina (2004), and then apply it to the expanded two-stage RP model. Let the expanded random variables representing
response in a simple random sample, $Y_i = \sum_{s=1}^{N} U_{is} y_s$ be defined by

$$\hat{Y}_i = (Y_{i1} \ Y_{i2} \ \cdots \ Y_{iN})'$$

where $Y_{iu} = U_{is}s_{ys}$ and suppose we wish to predict $P = L \hat{Y}$

where $L = c' \otimes 1'_{N}, \ c = (c_i)$, and where $c_i$ are arbitrary constants. The theorem states
conditions under which we can develop optimal estimators of $P$ based on $Y = ((Y_i))$

instead of $\hat{Y}$ without loss of information. To use the theorem, note that we can express
\[ \tilde{Y} = X\tilde{Y} + P_X\tilde{Y}, \] where \( X = I_N \otimes \frac{1}{N} J_N \), \( P_X = I_N \otimes P_N \) and that \( X\tilde{Y} = \left( I_N \otimes \frac{1}{N} I_N \right) Y \).

As a result, \( P = L X\tilde{Y} + L P_X\tilde{Y} \). The first term, which reduces to \( L X\tilde{Y} = c'Y \) is orthogonal to the second, and the expected value of the second term is zero. As a result, optimal predictors of \( P \) can be developed based solely on \( Y \) without loss of information.

We divide the random permutation vector into a sample and remainder, and subsequently develop the unbiased constraint.

We partition the random variables into the sample and remainder. To do so, let us define

\[
\tilde{C}' = \begin{pmatrix} \tilde{C}'_I \\ \tilde{C}'_H \end{pmatrix}, \quad \tilde{C} = \begin{pmatrix} N \otimes \tilde{C}'_I \\ N \otimes \tilde{C}'_H \end{pmatrix}, \quad \tilde{C}'_I = \begin{pmatrix} n C_1 \\ \vdots \\ N C_N \end{pmatrix}, \quad \tilde{C}'_H = \begin{pmatrix} 0 \\ \vdots \\ N \otimes C_s \end{pmatrix},
\]

where \( C_s = \left( I_{m_1} , 0'_{M_s - m_1} \right) \) and \( C_s^\prime = \left( 0_{m_1} , I'_{M_s - m_1} \right) \). Notice that

We consider three types of collapsed random variables. The first two enable the development of unbiased predictors for a PSU total, or mean when the samples of SSUs are selected with PPS sampling. The third type enables predictors of a PSU mean to be
developed under any second stage simple random sampling plan, but imposes a weaker unbiased requirement. In the first case the random variables are collapsed to PSU totals for the sample and the remainder. In the second case the random variables are collapsed to scaled PSU totals for the sample and the remainder using the cluster size as the scaling factor. In the third strategy, the random variables are collapsed to PSU means for the sample and the remaining SSUs.

In order to simplify the derivation of the predictors, we introduce some additional notation. We represent the sampling fraction for cluster $s$ as $f_s = \frac{m_s}{M_s}$, and the average sampling fraction as $\bar{f} = \frac{\bar{m}}{\bar{M}}$, where $\bar{m} = \frac{1}{N} \sum_{s=1}^{N} m_s$ and $\bar{M} = \frac{1}{N} \sum_{s=1}^{N} M_s$. Ratios of size for the sample, remainder, and cluster are given by $r_{sl} = \frac{m_s}{\bar{m}}$, $r_{sll} = \frac{M_s - m_s}{\bar{M} - \bar{m}}$, and $r_s = \frac{M_s}{\bar{M}}$, and are used to define the weighted cluster means $\mu^*_s = \frac{1}{N} \sum_{s=1}^{N} \mu^*_{sl}$, $\mu^*_{ll} = \frac{1}{N} \sum_{s=1}^{N} \mu^*_{sll}$, $\mu^*_R = \frac{1}{N} \sum_{s=1}^{N} \mu^*_{R}$, $\mu_r^* = \frac{1}{N} \sum_{s=1}^{N} \mu_r^*_{sll}$, and $\mu^*_R = \frac{1}{N} \sum_{s=1}^{N} \mu^*_{R}$. Similarly, we define average deviations from these means by $\bar{\beta}^*_i = \frac{1}{N} \sum_{s=1}^{N} \beta^*_{sl} \mu^*_{sll}$, $\bar{\beta}^*_{ll} = \frac{1}{N} \sum_{s=1}^{N} \beta^*_{sll} \mu^*_{sll}$, $\bar{\beta}^*_R = \frac{1}{N} \sum_{s=1}^{N} \beta^*_{R} \mu^*_{R}$, $\bar{\beta}^*_r = \frac{1}{N} \sum_{s=1}^{N} \beta^*_r \mu^*_{sll}$ and

$\beta^*_R = \mu^*_R - \mu_R$. Vectors of length $N \times 1$ containing these means are denoted by
\( \mathbf{\mu}_{i} = (\mu_{si}), \ \mathbf{\mu}_{ii} = (\mu_{si}'), \ \mathbf{\mu}_{i} = (\mu_{si}'), \ \mathbf{\mu}_{ii} = (\mu_{si}''), \ \mathbf{\beta}_{i} = (\beta_{si}), \ \mathbf{\beta}_{ii} = (\beta_{si}' '), \ \mathbf{\beta}_{i} = (\beta_{si}'), \ \mathbf{\beta}_{ii} = (\beta_{si}''), \)

\[ \mathbf{\beta}_{i}' = (\beta_{si}'')', \ \mathbf{\beta}_{ii}' = (\beta_{si}''), \] and \( \mathbf{\beta}_{R} = (\beta_{si}''). \)

Variance components are defined by \( \sigma_{i}^{2} = \frac{1}{N-1} \sum_{s=1}^{N} \beta_{si}^{2} ', \) \( \sigma_{ii}^{2} = \frac{1}{N-1} \sum_{s=1}^{N} \beta_{si}^{2} '', \) and \( \sigma_{i,ii}^{2} = \frac{1}{N-1} \sum_{s=1}^{N} \beta_{si}^{2} s', \)

\( \sigma_{il}^{2} = \frac{1}{N-1} \sum_{s=1}^{N} \beta_{si}^{2} s', \) and \( \sigma_{l,il}^{2} = \frac{1}{N-1} \sum_{s=1}^{N} \beta_{si}^{2} s'' . \)

Additional weighted average variance components for within cluster variances are given by \( \sigma_{e}^{2} = \frac{1}{N} \sum_{s=1}^{N} f_{s} (1-f_{s}) r_{s} \sigma_{s}^{2} , \)

\( \sigma_{e}^{2} = \frac{1}{N} \sum_{s=1}^{N} f_{s} (1-f_{s}) \sigma_{s}^{2} r_{s} , \) \( \sigma_{e}^{2} = \frac{1}{N} \sum_{s=1}^{N} \sigma_{s}^{2} r_{s} , \) \( \sigma_{e}^{2} = \frac{1}{N} \sum_{s=1}^{N} \sigma_{s}^{2} r_{s} , \) and \( \sigma_{e}^{2} = \frac{1}{N} \sum_{s=1}^{N} \sigma_{s}^{2} r_{s} . \)

We also define \( k' = \frac{\bar{m} \sigma_{s}^{2}}{\sigma_{e}^{2} + \bar{m} \sigma_{s}^{2} }, \) \( k' = \frac{\bar{m} \sigma_{s}^{2}}{\sigma_{e}^{2} + \bar{m} \sigma_{s}^{2} }, \) \( k' = \frac{\bar{m} \sigma_{s}^{2}}{\sigma_{e}^{2} + \bar{m} \sigma_{s}^{2} }, \) where \( \sigma_{l}^{2} = \sigma_{l}^{2} - \frac{\sigma_{e}^{2}}{M} , \)

\( \sigma_{s}^{2} = \sigma_{s}^{2} - \frac{\sigma_{s}^{2}}{M} \) and \( \sigma_{s}^{2} = \sigma_{s}^{2} - \frac{\sigma_{s}^{2}}{M} . \) The details of the development of predictors are given next.

3.1. Predictors of PSU Totals Based on Collapsing to Cluster Totals

We consider a vector of PSU totals for the sample and remainder by letting

\( \mathbf{Y} = \mathbf{C}_{s} \mathbf{Y} \) where \( \mathbf{C}_{s} = \left( \mathbf{I}_{N} \otimes (\mathbf{C}_{N} \mathbf{I}_{N})' \right), \mathbf{C}_{s} = \mathbf{I}_{N} \otimes (\mathbf{C}_{N} \mathbf{I}_{N})', \mathbf{C}_{s} = \mathbf{I}_{N} \otimes (\mathbf{C}_{N} \mathbf{I}_{N})', \mathbf{C}_{s} = \mathbf{I}_{N} \otimes (\mathbf{C}_{N} \mathbf{I}_{N})', \) and
The vector $Y^*$ has dimension $2N \times 1$ with the first $N$ random variables consisting of totals of the sample SSUs for each PSU, namely $Y_{il}^* = \sum_{s=1}^N U_{is} m_s \overline{Y}_{sl}$, where $\overline{Y}_{sl} = \frac{1}{m_s} \sum_{j=1}^{m_s} \tilde{Y}_{sj}$, $i = 1, \ldots, N$, and the second set of $N$ random variables consisting of totals of the remaining SSUs for each PSU, namely $Y_{ill}^* = \sum_{s=1}^N U_{is} (M_s - m_s) \overline{Y}_{sll}$, $i = 1, \ldots, N$ with $\overline{Y}_{sll} = \frac{1}{M_s - m_s} \sum_{j=m_s+1}^{M_s} \tilde{Y}_{sj}$. The model for $Y^*$ is given by

$$Y^* = X^* \begin{pmatrix} \mu_t^* \\ \mu_t^* \\ \mu_t^* \end{pmatrix} + Z^* \begin{pmatrix} B_t^* \\ B_t^* \\ B_t^* \end{pmatrix} + E^*$$

where $X^* = \overline{M} \begin{pmatrix} f \\ 0 \\ 1 - f \end{pmatrix} \otimes I_N$, $Z^* = \overline{M} \begin{pmatrix} f \\ 0 \\ 1 - f \end{pmatrix} \otimes I_N$, $B^* = (B_{ll}^*) = U\beta^*$ and $B^* = (B_{ll}^*) = U\beta^*$.

The fixed effects correspond to the expected PSU total for the potential sample SSUs, $\overline{m} \mu_t^*$, and the expected PSU total for the remaining SSUs, $(\overline{M} - \overline{m}) \mu_t^*$. Notice that these parameters are defined as averages over all PSUs in the population. The random effects correspond to deviations of PSU totals (for the sample and the remainder) about their respective expected values, i.e. $B_{il}^* = \sum_{i=1}^N U_{is} \beta_{sl}^*$ and $B_{ill}^* = \sum_{i=1}^N U_{is} \beta_{sll}^*$. 
Notice that we can express \( \mu_i^* = \mu + \beta_i^* \) and \( \mu_{ii}^* = \mu + \beta_{ii}^* \), which in matrix notation may be written as,

\[
\begin{pmatrix}
\mu_i^* \\
\mu_{ii}^*
\end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}
\begin{pmatrix} \mu \\ \beta_i^* \\ \beta_{ii}^* \end{pmatrix}.
\]

Neither \( \beta_i^* \) nor \( \beta_{ii}^* \) are necessarily equal to zero. The two additional parameters correspond to weighted averages of sample cluster effects, and weighted averages of remainder cluster effects. The variance of the collapsed random variables is

\[
\text{var}_{\xi_i \tilde{\xi}_i} (Y_i^*) = M^2 \left[ \frac{\bar{f}^2 \sigma_i^2}{\bar{f}(1-\bar{f})\sigma_{ii}^2} \right] \otimes P_N + M(1-\bar{f})\sigma_i^2 \left[ \begin{array}{c} 1 \\ -1 \\ -1 \end{array} \right] \otimes I_N.
\]

Next, we develop a predictor of the total for PSU \( i \) expressed as \( P_i^* = g'_i Y^* \) where \( g_i = 1_2 \otimes e_i \). We require the predictor to be a linear function of the random variables in the sample, to be unconditionally unbiased, and to minimize the expected value of the mean squared error. First, we partition the random variables into a vector for the sample and a vector for the remainder by letting \( (Y_i^{s*} \quad Y_i^{r*})' = (K_i' \quad K_{ii}') Y^* \) where

\[
K_i = \left( \begin{array}{c|l}
I_n & 0 \\
\hline
0 & (N-n)N_n
\end{array} \right) \quad \text{and} \quad K_{ii} = \left( \begin{array}{c|l}
0 & I_{(2N-n)n} \\
\hline
I_{(N-n)N_n} & 0
\end{array} \right).
\]

Consequently,

\[
E_{s_i \tilde{s}_i} (Y_i^*) = \bar{m}1_n \mu_i^* \quad \text{and} \quad E_{s_i \tilde{s}_i} (Y_{ii}^*) = \left( \begin{array}{c}
\bar{m}1_{N-n} \\
0
\end{array} \right) \left( \begin{array}{c}
0 \\
\bar{m} \quad \bar{M} - \bar{m} \end{array} \right) \left( \begin{array}{c}
\mu_i^* \\
\mu_{ii}^*
\end{array} \right)
\]

The total for PSU \( i \) is given by \( P_i^* = g'_i Y_i^* + g'_{ii} Y_{ii}^* \) where \( g'_i = e'_i, \quad g'_{ii} = (e'_{ii} \quad e'_i), \) and \( e'_i = (e'_i \quad e'_{ii}) \) with the dimensions of \( e_i \) and \( e_{ii} \) being respectively, \( n \times 1 \) and \( (N-n) \times 1 \). Collection of the study data will result in realizing the values of \( Y_i^* \).
We require the predictor of \( P_i^* \) to be an unbiased linear function of the sample data, \( \hat{P}_i = (g_i' + a') Y_i' \), such that \( E_{s_i,i} \left( \hat{P}_i - P_i^* \right) = 0 \). This unbiased constraint implies that \( a' E_{s_i,i} \left( Y_i' \right) - g_i' E_{s_i,i} \left( Y_{ii}' \right) = 0 \), a condition that can be expressed as

\[
\frac{1}{N} \sum_{s=1}^{N} \left( c - \frac{1}{f_s} \right) m_s \mu_s = 0
\]

for some constant \( c \). To satisfy the unbiased constraint, we require \( c = \frac{1}{f_s} \) for all \( s = 1,\ldots,N \). Thus, an unbiased predictor of a PSU total exists only when second stage sampling fractions are equal for all clusters, i.e. second stage sampling is PPS.

Thus, we assume PPS second stage sampling to develop an unbiased predictor of the PSU total that minimizes the expected value of the MSE. The common second stage sampling fraction is represented by \( f \). With PPS sampling, \( f = \overline{f} \) and \( r_s = r_{sl} = r_{illi} \) so that \( \mu_R = \mu_i = \mu_{ii} \) and \( \sigma_R^2 = \sigma_i^2 = \sigma_{ii}^2 \), and hence

\[
\text{var}_{s_i,i} \left( Y_i' \right) = \overline{M}^2 \sigma_R^2 \left( f \right) \left( 1 - f \right) \otimes P_N + \overline{M} f \left( 1 - f \right) \sigma_{c}^2 \left( \begin{array}{c} 1 \\ 1 \\ 1 \\ 1 \\ \end{array} \right) \otimes I_N.
\]

The model for the sample is \( Y_i' = X_i \mu_R + Z_i \beta_R + E_i' \) where \( X_i = \overline{M} f \mathbf{I}_n \),

\[
Z_i = \overline{M} \left( \begin{array}{cc} \mathbf{I}_n & \mathbf{0} \\ \end{array} \right)_{n \times \left( Y-n \right)} \text{ and } \beta_R = U \beta_R, \text{ while the model for the remainder is}
\]
\[ Y_{ii}^* = X_{ii}^* \mu_R + Z_{ii}^* B_R + E_{ii}^* \] where \( X_{ii}^* = \bar{M} \left( f^1_{N-n} \right) \) and \( Z_{ii}^* = \bar{M} \left( \begin{bmatrix} 0 \\ (N-n) f^1_{N-n} \end{bmatrix} \right) \).

Partitioning the variance accordingly, we get \( \text{var}_{\frac{N}{n}} (Y_i^* - Y_{ii}^*) = \begin{bmatrix} V_i^* \bar{V}_{i,ii}^* V_{i,ii}^* \end{bmatrix} \) (see Appendix).

The best linear unbiased predictor of the total for PSU \( i \) is given by
\[
\hat{P}_i = g_d' Y_i + g_{ill}' \left[ X_{i,ill}^* \mu_{ill} + V_{i,ii}^* \left( Y_i^* - X_i^* \mu_R \right) \right],
\]
where \( \mu_{ill} = \left( X_i^* V_{i,ii}^{-1} X_i^* \right)^{-1} X_i^* V_{i,ii}^{-1} Y_i^* \).

We express the total for PSU \( i \) as \( g_i' Y_i^* \) where \( g_i' = (g_d', g_{ill}) \), \( g_d' = e_d', \), \( g_{ill} = (e_{ill}', e_i') \), and \( e_i = (e_{ill}', e_i')' \). Simplifying terms it follows that \( \hat{\mu}_R = \bar{Y}_{RL} \) where \( \bar{Y}_{RL} = \frac{1}{n} \sum_{i=1}^{n} \bar{Y}_{ill} \) and
\[
\bar{Y}_{ill} = \sum_{s=1}^{N} U_{is} r_s \bar{Y}_{st}.
\]
Then, using \( \bar{Y}_{RL} = \left( \bar{Y}_{ill} \right) \), \( X_i^* = \bar{M} f^1_n \), \( X_{i,ill}^* = \bar{M} \left( f^1_{N-n} \right) \), \( g_d' = e_d' \), and \( g_{ill}' = (e_{ill}', e_i') \), when \( i \leq n \), we obtain
\[
\hat{P}_i = \bar{M} \left( f \bar{Y}_{ill} + \left( 1 - f \right) \left[ \bar{Y}_{RL} + k^* \left( \bar{Y}_{ill} - \bar{Y}_{RL} \right) \right] \right).
\]

When \( i > n \), it follows that \( \hat{P}_i = \bar{M} \bar{Y}_{RL} \).

Because of the unbiased condition, the MSE of the predictor is \( \text{var}_{\frac{N}{n}} \left( \hat{P}_i - P_i^* \right) \).

Using the expressions for the partitioned variance and simplifying, it follows that when
\[ i \leq n, \quad \text{var}_{\xi_i} \left( \hat{P}^*_i - P^*_i \right) = M^2 (1 - f) \left[ \frac{\sigma^2}{nm} + \left( \frac{n-1}{n} \right) (1 - k^*) \sigma^2_R \right] \] while \( i > n \),

\[ \text{var}_{\xi_i} \left( \hat{P}^*_i - P^*_i \right) = M^2 (1 - f) \left[ \frac{\sigma^2}{nm} + \left( \frac{n+1}{n} \right) \sigma^2_R \right]. \]

3.2. Predictors of PSU Means based on Weighted PSU Totals

Predictors of the PSU mean, \( P_i \), can be developed from a collapsed set of random variables corresponding to weighted PSU totals by letting \( Y^* = C^* Y \) with

\[
C^* = \left( I_N \otimes \left( C_1^* I_N \right) \right) \left( I_N \otimes \left( C^*_H 1_N \right) \right) \quad \text{and} \quad \text{var}_{C^*} = \left( I_N \otimes \left( C^*_1 1_N \right) \right),
\]

where \( C^*_i = \left( \frac{1}{M_s} \mathbf{y}_s - 0 \right) \) and \( C^*_H = \left( \frac{1}{M_s} \mathbf{y}_s (M_s - m_s) \right) \). The vector of random variables \( Y^* \) has dimension \( 2N \times 1 \), with the first \( N \) elements consisting of weighted totals of the sample SSUs for each PSU,

\[
\bar{y}_{ij}^* = \sum_{s=1}^{N} U_{is} f_i \bar{y}_{s}^*, \quad i = 1, \ldots, N,\]

and the second set of \( N \) elements consisting of weighted totals of the remaining SSUs for each PSU,

\[
\bar{y}_{ij}^* = \sum_{s=1}^{N} U_{is} (1 - f_i) \bar{y}_{s}^*, \quad i = 1, \ldots, N.
\]

The model for \( Y^* \) is given by

\[
Y^* = X^* \left( \begin{array}{c} \mu^*_i \\ \text{var}^*_i \\ \mu^*_H \\ \text{var}^*_H \end{array} \right) + Z^* \left( \begin{array}{c} B^*_i \\ B^*_H \end{array} \right) + E^*.
\]
where $X^* = \frac{1}{M} X^*$, $Z^* = \frac{1}{M} Z^*$, $B^*_i = \left( \left( B^*_i \right) \right) = U^*_i b$ and $B^*_ii = \left( \left( B^*_ii \right) \right) = U^*_ii$. The fixed effects can be reparameterized by writing the models for the PSU sample and remainder totals in terms of the population mean and weighted average deviations, such that

$$
\begin{pmatrix}
\mu^*_i \\
\mu^*_ii
\end{pmatrix} = 
\begin{pmatrix}
\bar{f}^* \\
\bar{f}^* (1 - \bar{f})
\end{pmatrix} \begin{pmatrix}
1 \\
0
\end{pmatrix} \begin{pmatrix}
\mu \\
\bar{\beta}^*_i
\end{pmatrix}.
$$

The corresponding variance is given by

$$
\text{var}_{\alpha^2}(Y^*) = 
\begin{pmatrix}
\bar{f}^* \sigma^2_i \\
\bar{f}^* (1 - \bar{f}) \sigma^2_{ii}
\end{pmatrix} \otimes P_N + \bar{f} (1 - \bar{f}) \frac{\sigma^2}{M} \left[ \begin{pmatrix}
1 \\
-1
\end{pmatrix} \otimes I_N \right].
$$

We develop a predictor of the mean for PSU $i$ defined by $P_i = g_i Y^*$ where $g_i = 1_2 \otimes e_i$. The development is similar to that in Section 3.1. We require the predictor to be a linear function of the random variables in the sample, to be unconditionally unbiased, and to minimize the expected value of the mean squared error. First, we partition the random variables into a vector for the sample and the remainder by setting

$$
\begin{pmatrix}
Y^*_i \\
Y^*_ii
\end{pmatrix} = (K_i' K_{ii}') Y^*,
$$

and express the mean for PSU $i$ as $P_i = g_i' Y_i^* + g_{ii}' Y_{ii}^*$. We require the predictor of $P_i$ to be an unbiased linear function of the sample data,

$$
\hat{P}_i = (g_i' + a') Y_i^*,
$$

such that $E_{\xi_{i}} \left( \hat{P}_i - P_i \right) = 0$. The unbiased constraint will be satisfied only when $f_s$ is constant for all $s = 1,...,N$, implying that an unbiased predictor of the mean for PSU $i$ exists only when the second stage sampling is PPS.
Assuming PPS sampling, we have $\mu^*_i = \mu^*_n = \mu_n$ and hence

\[
\mu^*_i = \mu^*_n = (\mu_i) = \mu \quad \text{and} \quad \mu^*_i = \mu^*_n = \mu.
\]

Defining $B = UB$ and $\beta = ((\beta_i)) = \mu - \mu_1$, the model for the sample simplifies to $Y^*_i = X^*_i \mu + Z^*_i B + E^*_i$ where $X^*_i = f^*_1$ and

\[
Z^*_i = \left( f \quad I_n \quad 0 \right)_{n \times N},
\]

while the model for the remainder is given by

\[
Y^*_n = X^*_n \mu + Z^*_n B + E^*_n \text{ where } X^*_n = \left( f^*_1 \quad (1 - f) \right)_{N \times N}
\]

and

\[
Z^*_n = \left( \begin{array}{c}
0 \\
A^*_{N-n} \\
l_1
\end{array} \right)_{n \times N},
\]

The corresponding variance simplifies to

\[
\text{var}_{i \in \xi} (Y^*) = \sigma^2 \left( \begin{array}{c}
f \\
1-f
\end{array} \right) \odot P + f (1-f) \frac{\sigma^2}{M} \left( \begin{array}{c}
1 \\
-1
\end{array} \right) \odot I_N
\]

since $\sigma^2_i = \sigma^2_n = \sigma^2_{i,n} = \sigma^2$. We partition the variance in a similar manner as in Section 3.1 to form the predictor.

The best linear unbiased predictor of a PSU mean when $i \leq n$ is

\[
\hat{P}^*_i = f \bar{Y}_i + (1-f) \left[ \bar{Y} + k^* (\bar{Y}_i - \bar{Y}) \right]
\]

and when $i > n$

\[
\hat{P}^*_i = \bar{Y}
\]

where $\bar{Y}_i = (\bar{Y})$, $\bar{Y}_i = \sum_{s=1}^N U_i \bar{Y}_s$ and $\bar{Y} = \frac{1}{n} \sum_{i=1}^n \bar{Y}_i$. The MSE of the predictor when $i \leq n$ is

\[
\text{var}_{i \in \xi} (\hat{P}^*_i - \bar{P}_i) = (1-f) \left[ \frac{\sigma^2}{mn} + \left( \frac{n-1}{n} \right) (1-k^*) \sigma^2 \right].
\]
while when $i > n$, 
\[
\text{var}_{\xi_i} \left( \hat{P}_i - P_i \right) = (1 - f) \left[ \frac{\sigma_{ij}^2}{n\bar{m}} + \frac{(n+1)}{n} \frac{\sigma_i^2}{1 - f} \right].
\]

3.3. Predictors of a PSU Mean based on Sampled and Remaining SSU Averages

Predictors of a PSU mean can be developed from a collapsed set of random variables corresponding to sampled and remaining SSU averages by setting \( \mathbf{C}^* \mathbf{Y} = \mathbf{Y}^* \)

where \( \mathbf{C}^* = \left( \mathbf{I}_N \otimes \left( \mathbf{C}_1 \mathbf{1}_N \right) \right) \left( \mathbf{1}_N \otimes \left( \mathbf{C}_0 \mathbf{1}_N \right) \right)' \), with \( \mathbf{C}_i = \sum_{s=1}^{N} \left( \frac{1}{m_i} \mathbf{1}_m' \ 0 \right)' \) and

\[
\mathbf{C}^*_i = \sum_{s=1}^{N} \left( \frac{1}{M_N - m_i} \mathbf{1}_M' \ (M_N - m_i) \right)'.
\]

This form of collapsing the relevant random variables enables best linear unbiased predictors to be developed for second stage sampling plans when sampling fractions differ between clusters. The \( 2N \times 1 \) vector of random variables \( \mathbf{Y}^* \) is such that its first \( N \) elements consist of PSU sample means, \( \bar{Y}_i \), \( i = 1, \ldots, N \), and the second set of \( N \) elements consist of means for the remaining SSUs in each PSU,

\[
\bar{Y}_{iS} = \sum_{s=1}^{N} U_{is} \bar{Y}_{iS}.
\]

Using this collapsed set of random variables, the model simplifies to

\[
\mathbf{Y}^* = \mathbf{X}^* \mu + \mathbf{Z}^* \mathbf{B} + \mathbf{E}^*
\]

where \( \mathbf{X}^* = \mathbf{1}_{2N} \) and \( \mathbf{Z}^* = \mathbf{1}_2 \otimes \mathbf{I}_N \), and \( \mathbf{E}^* = \mathbf{C}^* \mathbf{E} \). The variance is

\[
\text{var}_{\xi_i} \left( \mathbf{Y}^* \right) = \left( \begin{array}{ccc} \frac{\sigma_{ij}^2}{\bar{m}} & 0 & 0 \\ 0 & \frac{\sigma_{ij}^2}{\bar{M} - \bar{m}} & 0 \\ 0 & 0 & \frac{\sigma_{ij}^2}{\bar{M} - \bar{m}} \end{array} \right) \otimes \mathbf{I}_N + \sigma^2 \mathbf{J}_2 \otimes \mathbf{I}_N - \frac{\sigma_i^2}{N} \mathbf{J}_2 \otimes \mathbf{J}_N.
\]
In order to develop a predictor of the PSU mean, we define a linear combination of the random variables as

\[ P_i^* = g_i^* Y^* = \sum_{s=1}^{N_i} U_i \left( c_i \bar{Y}_{ii} + (1-c_i) \bar{Y}_{il} \right) \]

where \( g_i^* = (c_i \ 1-c_i) \otimes e_i' \) for some constant \( c_i \). Notice that \( c_i \) depends on the position of the PSU, but not on the realized cluster. This random variable, although not necessarily equal to the PSU mean, has an expected value (over SSUs) that equals the mean for PSU \( i \), i.e., \( E_{\bar{Y}_i} (P_i^*) = P_i \). We develop a predictor of \( P_i^* \), not \( P_i \).

Again, we partition \( Y^* \) into a sample and remaining vector by letting

\[
\begin{pmatrix}
K_i \\
K_{ii}
\end{pmatrix} Y^* = \begin{pmatrix}
Y_i^* \\
Y_{ii}^*
\end{pmatrix}
\]

and similarly, we partition \( g_i^* \) by setting \( g_i^* (K_i' \ K_{ii}') = (g_{ii}^* \ g_{ii}^*) \)

such that \( g_{ii}^* = c e_i' \) and \( g_{ii}^* = (c e_i' \ 1-c_i e_i' \) \). Collection of the study data will result in realizing the values of \( \bar{Y}_i \). We require the predictor of \( P_i^* \) to be a linear function of the sample data, \( \hat{P}_i^* = (g_{ii}^* + a') \bar{Y}_i \) and be unbiased so that \( E_{\bar{Y}_i} (\hat{P}_i^* - P_i^*) = 0 \). The constraint will be satisfied if \( a'1_{n} = 1 - c_i \) for \( i \leq n \), or if \( a'1_{n} = 1 \) for \( i > n \). The best linear unbiased predictor is

\[
\hat{P}_i^* = g_{ii}^* \bar{Y}_i + g_{ii}^* \left[ X_i^* \bar{Y} + V_i^* (\bar{Y}_i - \bar{Y}) \right]
\]

which simplifies to

\[
\hat{P}_i^* = c_i \bar{Y}_i + (1-c_i) \left[ \bar{Y} + k^* (\bar{Y}_i - \bar{Y}) \right]
\]

when \( i \leq n \) and to \( \hat{P}_i^* = \bar{Y} \) when \( i > n \). When \( i \leq n \), the expected MSE simplifies to
\[
\text{var}_{\hat{\xi}_{i}^{*}}(\hat{P}_{i}^{*} - P_{i}^{*}) = (1 - c_{i}) \left[ \frac{\sigma_{\hat{\xi}^{*}}^{2}}{\hat{m}n} + \frac{(n-1)}{n} (1 - k_{i}) \sigma^{2} \right] + (1 - c_{i}) \left[ \frac{1 - c_{i}}{1 - f_{i}} \frac{\sigma_{\hat{\xi}^{*}}^{2}}{\hat{M}} - \frac{\sigma_{\hat{\xi}^{*}}^{2} + c_{i} \hat{M} \sigma^{2}}{\sigma_{\hat{\xi}^{*}}^{2} + fM \sigma^{2}} \right] \]
\[
 - \frac{1 - c_{i}}{n} \left[ \frac{c_{i}}{f_{i}} - \frac{1}{\frac{\sigma_{\hat{\xi}^{*}}^{2} + c_{i} \hat{M} \sigma^{2}}{\sigma_{\hat{\xi}^{*}}^{2} + fM \sigma^{2}}} \right] \sigma_{\hat{\xi}^{*}}^{2} \frac{1}{\hat{M}}
\]

and when \( i > n \) it reduces to
\[
\text{var}_{\hat{\xi}_{i}^{*}}(\hat{P}_{i}^{*} - P_{i}^{*}) = \frac{\sigma_{\hat{\xi}^{*}}^{2} - f_{i} \sigma_{\hat{\xi}^{*}}^{2}}{nm} + \sigma^{2} \left( \frac{n+1}{n} \right) + \left[ c_{i} \left( \frac{c_{i}}{f_{i}} \right) \frac{\sigma_{\hat{\xi}^{*}}^{2}}{M} + (1 - c_{i}) \left( \frac{1 - c_{i}}{1 - f_{i}} \right) \frac{\sigma_{\hat{\xi}^{*}}^{2}}{M} - \frac{\sigma_{\hat{\xi}^{*}}^{2}}{M} \right].
\]

4. APPLICATION

We illustrate predictors for a selected physician’s practice in the randomized controlled trial described in the Introduction. Patients enrolled in the study had duplicate venous cholesterol measures made at baseline and at 1 year follow-up. In addition, a 24 hour dietary recall was reported at baseline and at the 1 year follow-up. Data from these measures, in addition to the baseline questionnaire data and intervention data were used for study evaluation.

We limit discussion to a single treatment arm (III) to which fourteen physician practices were randomly assigned from \( N = 60 \). A summary of the average change for different sample physician practices assigned to this intervention is given in Table 1.
We focus on predicting the average change in the percent of Kcal from saturated fat for a particular realized physician practice (ID=71) for which the observed average change was -1.47% Kcal for the 25 sample patients. This estimate of the latent value for practice #71 is derived from a model that represents physician practices as fixed effects. This estimator is the best linear unbiased estimator of the parameter for cluster #71. Many practitioners still use such an estimator, ignoring the first stage sampling of clusters.

Using a model that accounts for the two stage design, the parameter for a cluster corresponds to the latent value of a realized PSU. Predictors of the realized latent value
have smaller expected MSE than the sample cluster mean. We can represent the predictors as the sum of two terms, one for the sample and the other for the remainder. For comparison, we have represented the sample mean as a similar sum by including an arbitrary non-zero constant, $c_i^*$ (see Table 2).

---

Table 2. Predictors of latent values of a realized PSU Mean based on Different models

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Target</th>
<th>Sample SSUs</th>
<th>Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{Y}_i$</td>
<td>Simple Mean</td>
<td>$\mu_i$</td>
<td>$1 - c_i^<em>$ $\overline{Y}_i$ + $c_i^</em>$ $\overline{Y}_i$</td>
</tr>
<tr>
<td>$\hat{P}_{MM}$</td>
<td>Mixed Model</td>
<td>$P_i$</td>
<td>$\hat{P}_{MM}^2 k_i (\overline{Y} - \mu)$</td>
</tr>
<tr>
<td>$\hat{P}_{SS}$</td>
<td>Scott&amp;Smith</td>
<td>$P_i$</td>
<td>$f_i \overline{Y}<em>i$ + $(1 - f_i) \left[ \hat{P}</em>{SS}^2 + k_i (\overline{Y}_i - \mu^*) \right]$</td>
</tr>
<tr>
<td>$\hat{P}^*$</td>
<td>RP (PPS)</td>
<td>$P_i$</td>
<td>$f_i \overline{Y}_i$ + $(1 - f_i) \left[ \overline{Y} + k^* (\overline{Y}_i - \overline{Y}) \right]$</td>
</tr>
<tr>
<td>$\hat{P}^*$</td>
<td>RP (General)</td>
<td>$P_i^*$</td>
<td>$c_i \overline{Y}_i$ + $(1 - c_i) \left[ \overline{Y} + k^* (\overline{Y}_i - \overline{Y}) \right]$</td>
</tr>
</tbody>
</table>
The model underlying the mixed model predictor is posited for a conceptually very large clustered population; it does not account for possibly different size clusters. The superpopulation model and the random permutation model incorporate cluster and sample sizes, and require these characteristics to evaluate the predictors. When second stage samples are not selected with PPS sampling, the target random variable, \( P^* \), is defined in terms of a non-zero constant \( c_i \). This constant can be arbitrary and still

\[
E_{\psi_i}(P^*_i) = P_i.
\]

In the Watch II study, physician practice sizes differed, but the practice size was not recorded. We make two assumptions about population sizes to illustrate differences in the predictors, first assuming half the eligible patients in a practice are included in the sample, and second assuming all physician practices have \( M = 100 \) eligible patients.

The predictors of the latent value for a realized PSU require knowledge (or estimates) of the population variance components. They are based on different models. To evaluate a predictor, we replace cluster variance components by their unbiased estimates. We use the mixed model restricted maximum likelihood estimate of the variance between clusters, \( \hat{\sigma}^2 = 0.431 \), and the results are summarized in Table 3.
Table 3. Predictors of the Latent Value for Physician Practice #71 using Empirical Variances assuming $M = 100$ Eligible Patients in each Cluster.

PPS Sampling ($f = 0.5$)

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictor=</th>
<th>Sample SSUs</th>
<th>Remaining SSUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Mean</td>
<td>$-1.47 = c_i^* (-1.47)$</td>
<td>$(1 - c_i^*)[-1.47]$</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>$-0.96 = -0.69 + 0.35[-1.47 - (-0.69)]$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scott and Smith</td>
<td>$-1.22 = (0.5)(-1.47)$ + $(1 - 0.5)(-0.69 + 0.35[-1.47 - (-0.69)])$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rand. Perm.</td>
<td>$-1.19 = (0.5)(-1.47)$ + $(1 - 0.5)(-0.94 - 0.04[-1.47 - (-0.94)])$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unequal Sampling of SSUs ($M = 100$ for all PSUs, varying $f_i$)

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictor=</th>
<th>Sample SSUs</th>
<th>Remaining SSUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Mean</td>
<td>$-1.47 = -1.47$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Mixed \[= -0.96 = -0.69 + 0.35 \left[ -1.47 - (-0.69) \right] \]

Scott and Smith \[-1.09 = (0.25)(-1.47) + \]
\[= (1 - 0.25)(-0.69 + 0.35 \left[ -1.47 - (-0.69) \right]) \]

Rand. Perm.* \[-1.08 = (0.22)(-1.47) + \]
\[= (1 - 0.22)(-0.94 + 0.06 \left[ -1.47 - (-0.94) \right]) \]

* \(c_i\) is set equal to \(\bar{m}/\bar{M}\)

The differences in the predictors in each panel in Table 3 highlight the different model assumptions. Each of the predictors has lower expected mean squared error relative to the simple mean estimator when evaluated over the same probability framework. For example, with PPS sampling, the expected MSE for the sample mean and random permutation model are 1.48 and 0.42, respectively.

5. DISCUSSION
Predicting random effects based on a two-stage random permutation model for cluster sampling is complicated by unequal size clusters, necessitating careful tracking of target random variables through the development of predictors. Tracing through this development, while more complex, retains a clear connection between the target random variable and the latent value of a the PSU. Avoiding these complications with other methods requires additional assumptions and introduces ambiguity in the interpretation of predictors. Perhaps more importantly, the usual mixed model methods ignores the structure of the actual population, and thus prevents tailoring of methods that account for such structure.

We have illustrated that it is possible to directly account for the physical randomization in a two stage simple random sample of an unequal size clustered population. This provides a design-based framework for prediction similar to that used in equal size cluster populations by Stanek and Singer (2004). In so doing, expanded sets of random variables are introduced to avoid random dimensions of PSUs, and then collapsed to PUS random variables prior to prediction. This strategy has the potential to be extended to include additional cluster level covariables, similar to work of Li (2003). We anticipate that extensions to include unit level covariates will require further expansion of the random variables for SSUs.

The development highlights other aspects of mixed model and super-population model predictors that differ from those developed here. Mixed model predictors are functions of the cluster sample size and cluster variances conditional on the realized cluster. As a result, the predictors can not be motivated as linear predictors of the random
variables unconditionally in any of the collapsed settings. It may be possible to specify a
different class of predictors in a random permutation model framework that come closer
to the mixed model predictors. Additional work is being conducted in this direction.
APPENDIX: EVALUATION OF $\text{var}_{\xi_2} (Y)$, $\text{var}_{\xi_2} (Y^2)$

We outline the steps leading to the expression for $\text{var}_{\xi_2} (Y)$ using the conditional expansion $\text{var}_{\xi_2} (Y) = \text{var}_{\xi_2} \left( E_{\xi_2 | \xi_1} [Y] \right) + E_{\xi_2 | \xi_1} \left( \text{var}_{\xi_2 | \xi_1} [Y] \right)$. Recall that the vector

\[ Y = \begin{pmatrix} Y_1' & Y_2' & \cdots & Y_N' \end{pmatrix}' \]

is constructed from sub-vectors for each PSU given by

\[ Y_i = \left( U_{i1} Y_{1i}' \quad U_{i2} Y_{2i}' \quad \cdots \quad U_{iN} Y_{Ni}' \right)' \]

where $Y_{ji} = \left( \tilde{Y}_j \right) = U_j (^{(s)} y_s, \tilde{Y}_j = \sum_{j=1}^{M} U_j (^{(s)} y_{jt} \text{ and } U_j^{(s)}$ is an $M_s \times M_s$ matrix of indicator random variables, $U_j^{(s)} = \begin{pmatrix} U_{j1}^{(s)} & U_{j2}^{(s)} & \cdots & U_{jM_i}^{(s)} \end{pmatrix}'$. We begin by obtaining the expected value and variance of these sub-vectors.

First, since $E_{\xi_2 | \xi_1} \left( U_j^{(s)} \right) = \frac{1}{M_s} J_{M_s}$, it follows that $E_{\xi_2 | \xi_1} (Y_i) = \left( \bigoplus_{s=1}^{N} I_{M_s} , \mu_s \right) \begin{pmatrix} U_{i1} \\ U_{i2} \\ \vdots \\ U_{iN} \end{pmatrix}$.

Defining $U = \begin{pmatrix} U_1 & U_2 & \cdots & U_N \end{pmatrix}$ where $U_s = \begin{pmatrix} U_{1s} & U_{2s} & \cdots & U_{Ns} \end{pmatrix}'$ and combining terms, $E_{\xi_2 | \xi_1} (Y) = \left[ I_N \bigotimes \left( \bigoplus_{s=1}^{N} I_{M_s} , \mu_s \right) \right] \text{vec} (U)$. Now, we develop an expression for

\[
\text{var}_{\xi_1} \left( \text{vec} (U) \right) = E_{\xi_1} \left[ \begin{pmatrix} U_1 U_1' & U_1 U_2' & \cdots & U_1 U_N' \\ U_2 U_1' & U_2 U_2' & \cdots & U_2 U_N' \\ \vdots & \vdots & \ddots & \vdots \\ U_N U_1' & U_N U_2' & \cdots & U_N U_N' \end{pmatrix} - \frac{1}{N^2} J_N \bigotimes J_N \right]
\]
simple random without replacement sampling, \( E_{\xi} \left( U_s U'_s \right) = \frac{1}{N} I_N \), while when \( s \neq s' \),

\[
E_{\xi} \left( U_s U'_{s'} \right) = \frac{1}{N(N-1)} (J_N - I_N). 
\]
Substituting these terms and simplifying,

\[
\text{var}_{\xi} \left( \text{vec}(U) \right) = \frac{1}{N-1} P_N \otimes P_N, \quad \text{and hence}
\]

\[
\text{var}_{\xi} \left( E_{\xi[|Y]} \right) = \frac{1}{N-1} P_N \otimes \left[ \left( \bigoplus_{s=1}^{N} I_{M_s} \mu_s \right) P_N \left( \bigoplus_{s=1}^{N'} I'_{M_s} \mu_s \right) \right].
\]

To evaluate \( E_{\xi} \left( \text{var}_{\xi[|Y]} \right) \) we expand the terms in

\[
\text{var}_{\xi[|Y]} = \begin{pmatrix}
\text{var}_{s_1 s_1} (Y_1) & \text{var}_{s_1 s_2} (Y_1, Y_2) & \cdots & \text{var}_{s_1 s_N} (Y_1, Y_N) \\
\text{var}_{s_2 s_1} (Y_2, Y_1) & \text{var}_{s_2 s_2} (Y_2) & \cdots & \text{var}_{s_2 s_N} (Y_2, Y_N) \\
\vdots & \vdots & \ddots & \vdots \\
\text{var}_{s_N s_1} (Y_N, Y_1) & \text{var}_{s_N s_2} (Y_N, Y_2) & \cdots & \text{var}_{s_N s_N} (Y_N)
\end{pmatrix}
\]

Now

\[
\text{var}_{\xi[|Y]} = \begin{pmatrix}
U_{i_1} U_{r_1} \text{var}_{s_1 s_1} \left( \tilde{Y}_1 \right) & U_{i_1} U_{r_2} \text{var}_{s_1 s_2} \left( \tilde{Y}_1, \tilde{Y}'_2 \right) & \cdots & U_{i_1} U_{r_N} \text{var}_{s_1 s_N} \left( \tilde{Y}_1, \tilde{Y}'_N \right) \\
U_{i_2} U_{r_1} \text{var}_{s_2 s_1} \left( \tilde{Y}_2, \tilde{Y}'_1 \right) & U_{i_2} U_{r_2} \text{var}_{s_2 s_2} \left( \tilde{Y}_2 \right) & \cdots & U_{i_2} U_{r_N} \text{var}_{s_2 s_N} \left( \tilde{Y}_2, \tilde{Y}'_N \right) \\
\vdots & \vdots & \ddots & \vdots \\
U_{i_N} U_{r_1} \text{var}_{s_N s_1} \left( \tilde{Y}_N, \tilde{Y}'_1 \right) & U_{i_N} U_{r_2} \text{var}_{s_N s_2} \left( \tilde{Y}_N, \tilde{Y}'_2 \right) & \cdots & U_{i_N} U_{r_N} \text{var}_{s_N s_N} \left( \tilde{Y}_N \right)
\end{pmatrix}
\]

Furthermore, since permutations of SSUs are independent for different PSUs,

\[
\text{var}_{\xi} \left( \tilde{Y}_s, \tilde{Y}'_s \right) = \bigoplus_{s=1}^{N} \text{var}_{s_1 s_1} \left( \tilde{Y}_s \right). \quad \text{Since } \tilde{Y}_s = \text{vec}(\tilde{Y}'_s) = \left( I_{M_s} \otimes y'_s \right) \text{vec}(U^{(s)}) \text{, we find that}
\]

\[
\text{var}_{\xi} \left( \text{vec}(U^{(s)}) \right) = \frac{1}{M_s - 1} P_{M_s} \otimes P_{M_s}, \quad \text{and hence, } \text{var}_{\xi} \left( \tilde{Y}_s \right) = P_{M_s} \sigma^2_s \text{ where}
\]

\[
\sigma^2_s = \frac{1}{M_s - 1} y'_s P_{M_s} y_s. \quad \text{Using these expressions, we get}
\]
\[
\text{var}_{\xi \sim \xi} (Y) = \left( \bigoplus_{s=1}^{N} U_{1s} U_{1s} P_{M_s} \sigma^2_s \right) \cdots \left( \bigoplus_{s=1}^{N} U_{Ns} U_{Ns} P_{M_s} \sigma^2_s \right).
\]

Finally, evaluating the expected value over permutations of PSUs,

\[
E_{\xi} \left( \text{var}_{\xi \sim \xi} [Y] \right) = I_N \bigotimes \left( \bigoplus_{s=1}^{N} \frac{1}{N} P_{M_s} \sigma^2_s \right).
\]

Under PPS sampling,

\[
\text{var}_{\xi \sim \xi} (Y^*) = \bar{M}^2 \sigma^2_{\bar{M}} \left( f \left( \frac{f}{1-f} \right) \bigotimes P_N + \bar{M}f (1-f) \sigma_{\epsilon}^2 \left( \frac{1}{-1} \frac{1}{-1} \frac{1}{1} \right) \bigotimes I_N \right) \text{ and }
\]

\[
\text{var}_{\xi \sim \xi} \left( Y_{f}^{*} \quad Y_{II}^{*} \right) = \begin{pmatrix} V_{f}^{*} & V_{I,II}^{*} \\ V_{f}^{*} & V_{II}^{*} \end{pmatrix}, \text{ such that}
\]
$$\text{var}_{\text{sys}}(Y_i) = \begin{vmatrix}
\begin{array}{c}
\frac{f^2}{N} \left( I_n - \frac{1}{N} J_n \right)
\end{array}
\end{vmatrix}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
-\frac{f^2}{N} \frac{1}{N} J
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
\frac{f(1-f)}{N} \left( I_n - \frac{1}{N} J_n \right)
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
-\frac{f(1-f)}{N} \frac{1}{N} J
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
(1-f) P_N
\end{array}
\end{vmatrix}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
I_n
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
0
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
-I_n
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
0
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
0
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
-I_n
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
I_n
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
0
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
0
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
-I_n
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
I_n
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
0
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
0
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
-I_n
\end{array}
\end{vmatrix}
\begin{vmatrix}
\begin{array}{c}
I_n
\end{array}
\end{vmatrix}
$$
REFERENCES


