Performance of Balanced Two-Stage Empirical Predictors of Realized Cluster Latent Values from Finite Populations: A Simulation Study

Silvina San Martino, Julio da Motta Singer, Edward J. Stanek

Abstract

Prediction of random effects in clustered finite populations is important in many practical problems. Predictors are usually based on the classical mixed effects model (assuming a conceptual infinite population), on the finite super-population model proposed by Scott and Smith (1969, JASA, 64: 830-840) or on the random permutation model recently proposed by Stanek and Singer (2004, JASA, 99:1119-1130). When variances are known and within cluster variances are equal, the random permutation model based predictor has smaller mean squared error than corresponding predictors obtained under either the mixed effects or Scott and Smith’s models. As population variances are rarely known, we propose method of moment estimators to obtain empirical predictors and

1 Facultad de Ciencias Agrarias, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina. silvina@ime.usp.br

2 Departamento de Estatistica, Universidade de São Paulo, São Paulo, Brazil. jmsinger@ime.usp.br

3 Department of Public Health, University of Massachusetts, Amherst, MA, USA. stanek@schoolph.umass.edu
conduct a simulation study to evaluate their performance. We identify situations (depending on the size of the populations, sampling fractions and intra-class correlations) where the random permutation model based predictor has smaller mean squared error than its competitors. For different within cluster variances, the simulation results suggest that the performance of the random permutation model based predictor improves either as the cluster sampling fractions increase or as the number and size of clusters increase. Also, when both cluster and unit intra-class correlation coefficients are very high (e.g., 0.95 or more), the performance of predictors derived under the three models is very similar. Additionally, the results indicate that when variances are known, the random permutation model based predictor derived under an equal within variance assumption has the best performance, even when they are different.

*Keywords: Finite population; Two-stage sampling; Random permutation; Optimal estimation; Prediction.*
1. Introduction

Best linear unbiased predictors (BLUP) based on mixed effects model theory are typically used for predicting realized random effects, i.e., random effects associated to realized units, such as cluster means (Goldberger (1962), Henderson (1984), McLean, Sanders, and Stroup (1991), Robinson (1991), Stanek, Well, and Ockene (1999), McCulloch and Searle (2001)). An assumption of the modeling approach is that the sample data are selected from a conceptual infinite (or very large finite) population of interest. In many settings, the underlying study population is finite, but this feature is rarely incorporated in the analysis.

Although Searle, Casella and Mc Culloch (1992) or Cox and Solomon (2003) mention that finite populations arise quite rarely in applications, there are instances where clustered finite populations occur naturally as in educational, public health or sociological surveys. Children in families, classrooms in schools or physician practices in hospitals are typical examples of such clusters, usually selected by random sampling without replacement from some finite population. Typically, there is interest in the contribution of within and between cluster variability to overall variability and in the prediction of expected response for a realized random cluster.

When interest lies in prediction of the average expected response for a realized cluster (I think we should simply use “the latent value of the realized cluster” or “the realized cluster’s latent value” as in the title. This is equal to the expected response of the realized cluster. I think this is clearer than the wording you use here.) based on a two-stage sample from a clustered finite population, three approaches are generally
considered. The first is the usual mixed model, but it does not account for the finite nature of the population. Although Searle and Fawcett (1970) developed a rule for converting expectations of mean squares obtained under variance component infinite population models into expectations under finite population models, these rules have been seldom used, due, in part, to a lack of theoretical results and software. The second approach, suggested by Scott and Smith (1969) and extended by Bolfarine and Zacks (1992) to include response error, considers the finite nature of the population and bases the inference on a super-population model. This method has had limited application, in part due to its sensitivity to model miss-specification and to the artificial nature of the postulated super-population. The third, suggested by Stanek, Singer and Lencina (2003) in a simple random sampling setup and extended by Stanek and Singer (2004) to two-stage sampling with or without response error, considers a design based probability model induced by the sampling process. This method does not require restrictive assumptions (such as normality) so that it can be applied to a wide range of practical settings.

Under each model, the predictors of realized cluster latent values are formed as the sum of predictors of observed and unobserved units, differing with respect to the associated shrinkage factors. These shrinkage factors are functions of population variance components and for the last two models they depend on finite population characteristics such as cluster sizes. In certain situations, the predictors obtained under the three models can differ greatly.

As an example, suppose that an educational survey is conducted in a given high-school to evaluate the ability of second graders with respect to a certain subject by means of a test with scores ranging from 0 to 10. We assume that the student responses include response error. To control for teacher effects, a two stage random sample is obtained from...
the population of second grade students assigned to classrooms (each with 30 students).
Assume that a sample of 15 students is selected from a sample of classrooms in the school.
In addition to estimating the school response and variance components, there may be
interest in predicting classroom response. Suppose that the between classroom variability is
1.25, the within cluster variability is 2 and that the response error variability is 0.8 (i.e., the
cluster intra-class correlation is 0.38 and the unit intra-class correlation is 0.71). Based on
the sample data, we can compute the average response for students in selected classrooms
and use the classroom average to estimate the average response for the school. Suppose the
ture school average is 6.75, while for the classroom with teacher \( i \), the sample average is
5.2. Assuming that the response error model holds for all students, the average performance
for the classroom with teacher \( i \) is predicted to be 5.4, 5.3 and 5.9 respectively using the
mixed effects model, Scott and Smith’s model or the random permutation model predictors.
The 11\% observed difference between the predicted values obtained under the random
permutation model and Scott and Smith’s model may be meaningful in this type of study.
Consequently, an evaluation of the performance of the predictors derived under these three
models for a wide range of conditions may be very helpful for practical applications. The
mean squared error may be used to select the best predictor. (Can we give these MSE’s?
For example, if \( n=5 \) of \( N=10 \) classrooms are selected, the MSE is given by ...)
Under the assumptions that all variances are known and within cluster variances are
equal, Stanek and Singer (2004) show that predictors of realized \( \text{cluster latent value} \) under
the random permutation model with response error have smaller mean squared error than
the mixed effects and Scott and Smith’s models. When within cluster variances are not

\[ \text{Silvina&al2005v8b_ed.doc - 07/06/05} \]
equal (but are known), a comparison of the MSE between predictors is not yet developed analytically.

In practical situations, variances are rarely known and need to be estimated. In this context, we propose estimators for such variances and report simulation study results that evaluate the performance of predictors of realized cluster latent values, providing guidance for the choice of predictors.

In Section 2 we present a brief review of the models and specify the corresponding predictors of cluster latent values. We also propose empirical predictors based on estimated variance components. In Section 3 we describe technical details of the simulation study to compare the performance of these predictors for finite populations with different structures. Finally, in Sections 4 and 5 we show simulation results and conclusions, respectively.

2. Predictors of cluster latent values under different models

We consider a finite population defined by a listing of \( M \) units, indexed by \( t = 1, \ldots, M \) in each of \( N \) clusters, indexed by \( s = 1, \ldots, N \). Associated with unit \( t \) in cluster \( s \) is a fixed constant \( y_{st} \) called an individual parameter. We summarize these parameters in the vector \( \mathbf{y} = \left( y_1', y_2', \ldots, y_N' \right)' \), where \( \mathbf{y}_s = \left( y_{s1}, y_{s2}, \ldots, y_{sM} \right)' \), \( s = 1, \ldots, N \).
We define the latent value of cluster $s$ as $\mu_s = \frac{1}{M} \sum_{i=1}^{M} y_{si}$ and the corresponding variance as $\left( \frac{M-1}{M} \right) \sigma^2_s = \frac{1}{M} \sum_{i=1}^{M} (y_{si} - \mu_s)^2$ for $s = 1, \ldots, N$. Also we let $\sigma^2 = \frac{1}{N} \sum_{s=1}^{N} \sigma^2_s$ denote the average within cluster variance. Similarly, we define the population mean and the between cluster variance as $\mu = \frac{1}{N} \sum_{s=1}^{N} \mu_s$ and $\left( \frac{N-1}{N} \right) \sigma^2 = \frac{1}{N} \sum_{s=1}^{N} (\mu_s - \mu)^2$, respectively.

We assume that a two-stage simple random sample is to be selected (without replacement) from this population. At the first stage, a sample of $n$ clusters is selected and at a second stage, a sample of $m$ distinct units is selected from the $M$ elements in each selected cluster.

2.1. The random permutation (RP) model

We define a probability model induced by the two-stage random sampling process that links the population parameters to an expanded vector of random variables. The two-stage random permutation model is represented as an ordered list of $NM$ random variables, the values of which are the response of an independent permutation of clusters and units in clusters. For each permutation, we assign a new label, $i = 1, \ldots, N$ to the clusters according to its position in the permuted list. In a similar manner, we label the positions in the permutation of units in a cluster by $j = 1, \ldots, M$. For ease of exposition, we refer to the cluster that will occupy position $i$ in the permutation of clusters as the $i^{th}$ primary sampling unit.
unit (PSU) and to the unit that will occupy position \( j \) in the permutation of units within a
cluster as the \( j^{th} \) secondary sampling unit (SSU). Since any unit in any cluster may occupy
position \( ij \), we represent the expected response for SSU \( j \) in PSU \( i \) as the random variable
\( Y_{ij} \).

To relate \( y_{ij} \) to \( y_{ij} \) we use two indicator random variables: \( U_{is} \), which takes on a
value of one when the realized cluster corresponding to PSU \( i \) is cluster \( s \) and a value of
zero otherwise, and \( U^{(*)}_{is} \), which takes on a value of one when the realized unit
corresponding to SSU \( j \) in cluster \( s \) is unit \( t \) and zero otherwise. As a consequence, the
random variable corresponding to SSU \( j \) in PSU \( i \) in a permutation is given by

\[
Y_{ij} = \sum_{i=1}^{N} \sum_{j=1}^{M} U_{is} U^{(*)}_{ij} y_{st}.
\]

The finite population \( y \) can be viewed as the realization of the random variable

\[
Y = \left(U \otimes I_{M}\right) \left(\bigoplus_{i=1}^{N} U^{(*)}_{i}\right) \mathbf{y},
\]

where \( \mathbf{Y} = \left(\mathbf{y}_{ij}\right) = \left(\mathbf{Y}_1' \quad \mathbf{Y}_2' \quad \cdots \quad \mathbf{Y}_N'\right) \in \mathbb{R}^{NM} \), with \( \mathbf{Y}_i = \left(Y_{i1} \quad Y_{i2} \quad \cdots \quad Y_{iM}\right)' \in \mathbb{R}^M \),

\( U^{(*)} = \left(\mathbf{U}_1^{(*)} \quad \mathbf{U}_2^{(*)} \quad \cdots \quad \mathbf{U}_M^{(*)}\right) \in \mathbb{R}^{M \times M} \), with \( \mathbf{U}_i^{(*)} = \left(U_{i1}^{(*)} \quad U_{i2}^{(*)} \quad \cdots \quad U_{iM}^{(*)}\right)' \), and

\( U = \left(\mathbf{U}_1 \quad \mathbf{U}_2 \quad \cdots \quad \mathbf{U}_N\right) \in \mathbb{R}^{N \times N} \), with columns \( \mathbf{U}_i = \left(U_{i1} \quad U_{i2} \quad \cdots \quad U_{iN}\right)' \).

Here, \( \otimes \) denotes the Kronecker product and \( \bigoplus_{i=1}^{N} \mathbf{A}_i \) denotes a block diagonal matrix with

blocks \( \mathbf{A}_i \) (Searle, Casella and McCulloch, 1992).
Defining $\beta_s = (\mu_s - \mu)$ as the deviation of the latent value of cluster $s$ from the population mean and $\varepsilon_{st} = (y_{st} - \mu)$ as the deviation of the individual parameter for unit $t$ (in cluster $s$) from the mean for cluster $s$, we can re-parameterize the vector of fixed values $y$ via the non-stochastic model

$$y = X\mu + Z\beta + \varepsilon$$  \hspace{1cm} (2.1)

where $X = I_N \otimes I_M$, $Z = I_N \otimes I_M$, $\beta' = (\beta_1, \beta_2, \ldots, \beta_N)$, and $\varepsilon$ is defined similarly to $y$.

Here, $I_a$ denotes an $a \times 1$ column vector with all elements equal to 1.

Using elementary properties of the indicator random variables and the simple structure of $X$ and $Z$, it follows that $U_1N = I_N$, $U^{(i)}_N I_M = I_M$, $(U \otimes I_M)\sum_{i=1}^{N} U^{(i)} = X = X$ and $(U \otimes I_M)\sum_{i=1}^{N} U^{(i)} = Z = U \otimes I_M = ZU$. Then, pre-multiplying both members of (2.1) by $(U \otimes I_M)\sum_{i=1}^{N} U^{(i)}$ and using the above results, we obtain the random permutation mixed effects model

$$Y = X\mu + ZB + E$$  \hspace{1cm} (2.2)

where $E = (U \otimes I_M)\sum_{i=1}^{N} U^{(i)}$, $B = U\beta = (B_1, B_2, \ldots, B_N)'$. Note that because of the random variables $U$, the terms $B_i = \sum_{i=1}^{N} U_{ii} \beta_j$ for $i = 1, \ldots, N$ are random effects and represent the deviation of the latent value for PSU $i$ from the population mean.

For the random variable $Y$ in (2.2) we have

$$E_{\tilde{Y}_{i|s}}(Y) = X\mu,$$
and

\[
\text{var}_{\xi_1 \xi_2} (\mathbf{Y}) = \sigma^2_e I_{NM} + \sigma^2 \left( I_N \otimes \mathbf{J}_N \right) - \frac{\sigma^2}{N} \mathbf{J}_{NM}
\]

where \( \sigma^2 = \sigma^2_e - \frac{\sigma^2}{M} \), \( \mathbf{J}_N = \mathbf{1}_N \mathbf{1}_N' \), and the subscripts \( \xi_1 \) and \( \xi_2 \) denote expectation with respect to permutations of the clusters and to permutations of units in a cluster, respectively.

As the values of the finite population defined by \( \mathbf{y} \) may not be observed directly, we assume a response error model of the form

\[
\tilde{\mathbf{Y}} = \mathbf{y} + \mathbf{W}
\]

where \( \mathbf{W} = (W_{st}) \in \mathbb{R}^{NM} \) is a vector of independent response errors with \( E(W_{st}) = 0 \) and \( \text{Var}(W_{st}) = \sigma^2_{st} \), \( s = 1, \ldots, N \), \( t = 1, \ldots, M \).

Using this notation, the two-stage random permutation model with response error is represented by

\[
\mathbf{Y}' = \left( \mathbf{U} \otimes \mathbf{I}_N \right) \left( \oplus_{s=1}^N \mathbf{U}^{(s)} \right) \tilde{\mathbf{Y}} = \mathbf{Y} + \mathbf{W}'
\]

where \( \mathbf{W}' = \left( \mathbf{U} \otimes \mathbf{I}_M \right) \left( \oplus_{s=1}^N \mathbf{U}^{(s)} \right) \mathbf{W} \). Under the re-parameterization (2.1), we can express this as a mixed effects model of the form

\[
\mathbf{Y}' = \mathbf{X}\mu + \mathbf{Z}\beta + \left( \mathbf{E} + \mathbf{W}' \right).
\]

(2.3)

The first two central moments of \( \mathbf{Y}' \) are

\[
E_{\xi_1 \xi_2 \xi_3} (\mathbf{Y}') = \mathbf{X}\mu
\]
and
\[
\text{var}_{\xi_3}(Y^*_{ij}) = \left( \sigma^2 + \sigma^2_{\xi} \right) I_{NM} + \sigma^2_{\xi,ij} (I_n \otimes J_m) - \frac{\sigma^2_{\xi}}{N} J_{NM}.
\]

The subscript $\xi_3$ denotes expectation with respect to response error and $\sigma^2 = \sum_{s=1}^{S} \sum_{t=1}^{T} \sigma^2_{st}$ denotes the average response error variance.

Note that each realization of the random variable $Y^*$ will generate the same finite population with (two-stage) permuted elements. Once the sample (i.e., permutation) has been selected, it will be apparent which cluster corresponds to a particular PSU. We refer to that cluster as the **realized PSU** and to the average of the expected response for the units of that cluster as the **latent value** for the realized PSU (noting that such effects are often defined as deviations relative to an overall mean). **When there is no response error, the latent value and the cluster mean coincide.**

Denoting the sample elements by $Y^*_i = Y_i + W^*_i$, the model for the sample is

\[
Y^*_i = X_i \mu + Z_i B + (E_i + W^*_i)
\]

where $X_i = I_n \otimes 1_m$, $Z_i = I_n \otimes 1_m$, $E(Y^*_i) = X_i \mu$ and

\[
\text{var}(Y^*_i) = \left( \sigma^2 + \sigma^2_{\xi} \right) (I_n \otimes I_m) + \sigma^2_{\xi,ij} (I_n \otimes J_m) - \frac{\sigma^2_{\xi}}{N} (J_n \otimes J_m).
\]

**2.2. Scott and Smith´s (SS) super-population model**

Scott and Smith (1969) used a super-population model according to which the finite population $y$ is viewed as a realization of a vector of random variables $Y$ such that
\[ E(\mathbf{Y}) = X\mu \quad \text{and} \quad \text{var}(\mathbf{Y}) = \sum_{i=1}^{N} \left( \sigma_i^2 \mathbf{I}_M + \sigma^2 \mathbf{J}_M \right) \] (2.4)

Note that the vector \( \mathbf{Y} \) in (2.4) is not directly linked to the population units as in the random permutation model. According to this model, elements within the same cluster are correlated, but elements in different clusters are not. In this context, \( \sigma^2 \) corresponds to the variance of cluster means. *I think it would be helpful to add some statements about what the interpretation is of \( \sigma_i^2 \). Suppose we let \( E_{cu} \) represent expectation with respect to sampling (from the super-population) of clusters (c) and units (u). Can we express the expected value and variance using these subscripts? I believe the usual interpretation of \( \sigma_i^2 \) is the variance of units in the cluster that was realized for the \( i \)th selection. This makes \( \sigma_i^2 \) a random variable, and only fixed if the selection is fixed. Thus,

\[
\text{var}_{cu} \left( Y_{ij} \mid i = s \right) = E_c \left[ \text{var}_{dc} \left( Y_{ij} \mid i = s \right) \right] + \text{var} \left( E_{du} \left[ Y_{ij} \mid i = s \right] \right) \\
= E_c \left[ \sigma_i^2 \mid i = s \right] + \text{var} \left( \mu_i \mid i = s \right)
\]

Now presumably \( E_c \left[ \sigma_i^2 \mid i = s \right] = \sigma^2 \). However, \( \text{var} \left( \mu_i \mid i = s \right) = 0 \), so there is problem with this interpretation, since it doesn’t result in the compound symmetric variance. So it is somewhat difficult to explain the interpretation. I think a long discussion of this will be distracting. However, some comment about difficulty in interpretation would help to notify the reader of such a problem.*

Using Bayesian models, Bolfarine and Zacks (1992) extend the approach considered by Scott and Smith (1969) to a two-stage sampling process with response error. Essentially, they add a normality assumption to the super-population model considered by
Scott and Smith (1969) and suppose that the potentially observed variables are given by elements of

\[ Y^* = Y + W, \]

where \( W \sim N\left(0, \sigma_r^2 I_M\right), \ i = 1, \ldots, N, \ j = 1, \ldots, M, \) and \( W \) is independent of \( Y \). Setting \( \sigma_r^2 = \sigma_r^2, \ i = 1, \ldots, N, \) we obtain

\[
E(Y^*) = X\mu \text{ and } \text{var}(Y^*) = \sum_{i=1}^{N} \left[(\sigma_r^2 + \sigma_r^2)I_M + \sigma^2 J_M\right]. \tag{2.5}
\]

Denoting the sample elements by \( Y_i^* = Y_i + W_i, \) it follows that \( E(Y_i^*) = X_i\mu \) and

\[
\text{var}(Y_i^*) = \sum_{i=1}^{n} \left[(\sigma_r^2 + \sigma_r^2)I_m + \sigma^2 J_m\right].
\]

2.3. The mixed effects (ME) model

Under a mixed effects model, the two-stage sample data are considered to have been selected from a conceptual infinite population, understood to be the limit (as the size becomes infinite) of the finite population of interest. In this case, the sample elements in \( \text{In my printout, the subscripts are sometimes bold, and sometimes not bold. They should not be bold.} \) PSU \( i, \) given by \( Y_i = (Y_{i1} Y_{i2} \ldots Y_{im})', \ i = 1, \ldots, n, \) may be modeled by

\[
Y_i = X_i\mu + Z_iB_i + E_i, \tag{2.6}
\]

where \( X_i = Z_i = I_n \) and \( E_i = (E_{i1} E_{i2} \ldots E_{im})'. \) Here, \( Y_{ij} \) is the response of SSU \( j, \ j = 1, \ldots, m \) in PSU \( i, \ i = 1, \ldots, n, \) \( \mu \) corresponds to the expected response over SSUs and...
PSUs in the conceptual infinite population, \( B_i \) is a random effect that corresponds to the deviation of the average expected response of SSUs in PSU \( i \) from \( \mu \), and \( E_{ij} \) is a random deviation of the expected response of SSU \( j \) from the average expected response of elements in PSU \( i \). Typically, it is assumed that \( B_i \sim N(0, \sigma^2_B) \) and \( E_{ij} \sim N(0, \sigma^2_{ij}) \) are independent, so that \( Y_{it} = \mu + B_i + E_{ij} \). If in (2.6) we include a second additive (response) error term \( W_{ij} \) then

\[
Y_{it}^* = \left( Y_{i1}^* \ Y_{i2}^* \ \cdots \ \ Y_{in}^* \right) = Y_{it} + W_{it} = \mu + \sigma_Y^{-1}(\sigma^2_B + \sigma^2_E)I_n + \sigma^2_J. 
\]

It follows that

\[
Y_{it}^* = X_{it}\mu + Z_{it}B_i + E_t + W_t, \tag{2.7}
\]

where \( B_i = (B_1 \ B_2 \ \cdots \ B_n) \), \( E_t = (E_{i1}' \ E_{i2}' \ \cdots \ E_{in}' \) \), and

\[
W_t = (W_{i1}' \ W_{i2}' \ \cdots \ W_{in}') \text{ which implies that } Y_{it}^* \sim N(X_{it}\mu + \sigma_Y^{-1}(\sigma^2_B + \sigma^2_E)I_n + \sigma^2_J). 
\]

The variance of the random effects, \( \sigma^2 \), can also be interpreted as the variance of the random cluster mean \( (\mu + B_i) \) which conceptually takes on an infinite number of values.

### 2.4. Predictors of the latent value of a realized PSU

Our principal interest lies in the linear combination that defines the latent value of PSU \( i \) (for \( i \leq n \)), i.e., \( T_i = g'Y \), where \( g' = \frac{1}{M}e_j \otimes I_M' \), and \( e_j \) denotes an \( N \times 1 \) column vector.
vector with 1 in position \( i \) and zero elsewhere. In the mixed model, \( T_i \) corresponds to
\[
\mu + B_i.
\]
The predictors of \( T_i \) (for \( i \leq n \)) under the three models (i.e. (2.3), (2.5), and (2.7))
are:

i) Mixed model: 
\[
\hat{T}_i = \hat{\mu} + k_i \left( \bar{Y}_i - \hat{\mu} \right) \quad \text{where} \quad \hat{\mu} = \frac{\sum_{i=1}^{n} \bar{Y}_i / v_i}{\sum_{i=1}^{n} 1 / v_i}, \quad \bar{Y}_i = \frac{1}{m} \sum_{j=1}^{m} Y_{ij},
\]
\[
v_i = m \sigma^2 + \sigma_i^2 + \sigma_r^2 \quad \text{and} \quad k_i = \frac{m \sigma^2}{m \sigma^2 + \sigma_i^2 + \sigma_r^2}.
\]

ii) Scott and Smith model: 
\[
\hat{T}_i = f \bar{Y}_i + (1 - f) \left[ \hat{\mu} + k_i \left( \bar{Y}_i - \hat{\mu} \right) \right] \quad \text{where} \quad \hat{\mu} \quad \text{and} \quad k_i \quad \text{are as in i) and} \quad f = \frac{m}{M} \quad \text{is the sampling fraction for units.}
\]

iii) Random permutation model: 
\[
\hat{T}_i = f \left[ \bar{Y} + k'_i \left( \bar{Y}_i - \bar{Y} \right) \right] + (1 - f) \left[ \bar{Y} + k^* \left( \bar{Y}_i - \bar{Y} \right) \right]
\]
where 
\[
k'_i = \frac{m \sigma^2 + \sigma_i^2}{m \sigma^2 + \sigma_i^2 + \sigma_r^2} \quad \text{and} \quad k^* = \frac{m \sigma^2}{m \sigma^2 + \left( \sigma_i^2 + \sigma_r^2 \right)} \quad \text{is defined earlier, but we need to define here the overall sample average,} \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} \bar{Y}_i.
\]

When no response error is present \( k'_i = 1 \), so that the predictor simplifies to
\[
\hat{T}_i = f \bar{Y}_i + (1 - f) \left[ \bar{Y} + k^* \left( \bar{Y}_i - \bar{Y} \right) \right].
\]

Note that \( k_i \) and \( k'_i \) are always positive, but \( k^* \) may be negative with minimum value
\[
- \frac{m}{M - m}.
\]
2.5. Empirical predictors of the latent value of a realized PSU

In practice, variance components are usually unknown and estimates are needed for \( k_i, k^*, \) and \( k'_r \). Empirical predictors can be obtained substituting \( k_i, k^*, \) and \( k'_r \) by their respective estimators. We estimate variance components under the limited assumptions of two stage sampling. When there is no response error, these estimators may be derived from ANOVA mean squares, namely

\[
MSB = (n-1) \sum_{i=1}^{n} \sum_{j=1}^{m} (\bar{y}_i - \bar{y})^2
\]  

and

\[
MSE = [n(m-1)]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \bar{y})^2.
\]

These mean squares may be expressed as quadratic forms of the type \( Y_i' A Y_j \) where

\[
A = (n-1)^{-1} \left( P_a \otimes \frac{J_m}{m} \right)
\]

for \( MSB \) or

\[
A = [n(m-1)]^{-1} \left( I_n \otimes P_m \right)
\]

for \( MSE \), with \( P_a = I_a - a^{-1}J_a \) and \( a \) denoting a positive integer.

Since \( A I_{m_e} = 0 \) and \( E(Y_i) = I_{m_e} \mu \) under any of the three models, it follows that

\[
E(Y_i' A Y_j) = 0.
\]

Therefore,

\[
E(Y_i' A Y_j) = tr \left[ var(Y_i) A \right] + E(Y_i' A E(Y_i)) = tr \left[ var(Y_i) A \right].
\]

Using these results, we may obtain method of moments estimators for the variance components, derive estimators for \( k_i, k^* \) and \( k'_r \) and obtain the corresponding empirical predictors as follows.

2.5.1. RP model
Under the random permutation model, we have
\[ \text{var}(Y_i) = \left( \sigma^2 + \sigma_r^2 \right) I_m + \sigma_r^2 \left( I_n \otimes J_m \right) - \frac{\sigma_r^2}{N} J_m. \]

To evaluate the expected value of MSE, we let \( A = [n(m-1)]^{-1} \left( I_n \otimes P_m \right) \), so that
\[ \text{var}(Y_i) A = \left( \frac{\sigma^2 + \sigma_r^2}{n(m-1)} \right) \left( I_n \otimes P_m \right), \]
which implies that
\[ E(MSE) = \sigma^2 + \sigma_r^2. \]

To evaluate the expected value of MSB, we let \( A = (n-1)^{-1} \left( P_n \otimes \frac{J_m}{m} \right) \), so that
\[ \text{var}(Y_i) A = \frac{\sigma^2 + \sigma_r^2}{n-1} \left( P_n \otimes \frac{J_m}{m} \right) + \frac{\sigma_r^2}{n-1} P_n \otimes J_m, \]
which implies that
\[ E(MSB) = \sigma^2 + \sigma_r^2 + m\sigma_r^2. \]

Assuming that the response error variance \( \sigma_r^2 \) is known and equating the observed and expected mean squares, we obtain \( \hat{\sigma}_r^2 + \sigma_r^2 = MSE \) and \( m\hat{\sigma}_r^2 = MSB - MSE \).

Consequently, the method of moments estimators for \( k^* \) and \( k_r^* \) are respectively given by
\[ \hat{k}^* = \frac{m\hat{\sigma}_r^2}{m\hat{\sigma}_r^2 + \hat{\sigma}_r^2 + \sigma_r^2} = \frac{MSB - MSE}{MSB} \quad \text{and} \quad \hat{k}_r^* = \frac{m\hat{\sigma}_r^2 + \hat{\sigma}_r^2}{m\hat{\sigma}_r^2 + \hat{\sigma}_r^2 + \sigma_r^2} = \frac{MSB - \sigma_r^2}{MSB}. \]
Since the minimum value of \( k^* \) is \( -\frac{m}{M-m} \) and \( \hat{k}_r^* \) must be positive, we propose the estimators
\[ \hat{k}^* = \max \left( -\frac{m}{M-m}, \frac{MSB - MSE}{MSB} \right), \]
and

\[ \hat{k}^* = \max \left( 0, \frac{MSB - \sigma^2}{MSB} \right) \]

to obtain the empirical predictors. [I am still curious of the performance of an estimator

where \( \hat{k}^* - \tilde{k}^* = \max \left( 0, \frac{MSE - \sigma^2}{MSB} \right) \)

2.5.2. ME and SS Models

Under the ME and SS models, we have

\[ \text{var} \left( \mathbf{Y}_j \right) = \bigoplus_{i=1}^{n} \left( a^2 \mathbf{I}_m + b^2 \mathbf{J}_m \right), \]

where \( a^2 = \sigma^2 + \sigma_i^2 \), and \( b^2 = \sigma^2 \). Using \( \mathbf{A} = [n(m-1)]^{-1} \left( \mathbf{I}_n \otimes \mathbf{P}_m \right) \), it follows that

\[ \text{var} \left( \mathbf{Y}_j \right) \mathbf{A} = \frac{1}{n(m-1)} \bigoplus_{i=1}^{n} a^2 \mathbf{P}_m, \]

and hence

\[ E(MSE) = \frac{1}{n} \sum_{i=1}^{n} a_i^2. \]

[This expression is somewhat confusing since it is not clear whether E(MSE) is a random
variable or a constant. I believe SS and the MM view it as a constant. If it is a constant,
then the \( a_i \) are a property of the 'position', not the cluster. We can imagine such a setting.

Suppose that the variance of units in a cluster is identical for all clusters, such that

\( \sigma_i^2 = \sigma_s^2 \) for all \( s \). Now suppose that different 'instruments' can be used to measure
response on a unit. Each instrument will result in the exact same response for the exact same unit. For example, suppose that the 'response' is the average score to a battery of questions. The different instruments use different sets of questions. A subject responds the exact same way to each question, but calculating the average response using the different instruments will result in a different 'response' for the same subject. In addition, if all subjects are evaluated in cluster “s” using instrument “i”, the variance in their response is given by \( \sigma_{s_i}^2 = \sigma_{\epsilon_i}^2 \). For simplicity, we might represent this variance as \( \sigma_{s_i}^2 = \sigma_{\epsilon_i}^2 = \sigma_i^2 \).

Thus, the variance using instrument “i” is not random.

Suppose now that we always use instrument 1 for the first selected cluster, instrument 2 for the second selected cluster, etc. Then E(MSE) will not be random. However, this setting has a different interpretation from the setting where the variance of units is different between clusters (with all measures based on the same instrument).

Letting \( \mathbf{A} = (n-1)^{-1} \left( \mathbf{P}_n \otimes \frac{\mathbf{J}_m}{m} \right) \), it follows that

\[
\text{var} (\mathbf{Y}_i) = \left[ m(n-1) \right]^{-1} \left[ \sum_{i=1}^{n} \left( \mathbf{P}_n \otimes \left( \mathbf{a}_i^2 + mb^2 \right) \right) \right] \otimes \mathbf{J}_m , \text{ and hence}
\]

\[
E(\text{MSB}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{a}_i^2 + mb^2 \tag{2.8}
\]

or equivalently,

\[
E(\text{MSB}) = \frac{1}{n} \sum_{i=1}^{n} \left( \mathbf{a}_i^2 + mb^2 \right) \tag{2.9}
\]

To obtain the empirical predictors, we will distinguish the following two cases:
a) Identical within cluster variances

When $\sigma^2 = \sigma^2$ for all $i = 1,\ldots,n$, we have $\sigma^2_i = \sigma^2 + \sigma^2_r$ for all $i = 1,\ldots,n$, and the expected mean squares terms reduce to

$$E(MSE) = \sigma^2 + \sigma^2_r$$

and

$$E(MSB) = \sigma^2 + \sigma^2_r + m\sigma^2.$$

Equating observed and expected mean squares, we obtain

$$\hat{k}_i = \max \left( 0, \frac{m\hat{\sigma}^2}{m\hat{\sigma}^2 + \hat{\sigma}^2_r + \sigma^2_r} \right) = \max \left( 0, \frac{MSB - MSE}{MSB} \right).$$

In this case, $\bar{\mu} = \bar{Y} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij}$ which is combined with $\hat{k}_i$ to form the predictors.

b) Different within cluster variances

We propose four estimators for the shrinkage constant $k_i = \frac{mb^2}{mb^2 + a^2_i}$, namely

$$i) \quad \hat{k}_{i(1)} = \begin{cases} 0 & \text{if } \frac{nm(\bar{Y}_i - \bar{Y})^2}{n-1} = 0 \\ \max \left( 0, 1 - \frac{S_i^2}{nm(\bar{Y}_i - \bar{Y})^2 / (n-1)} \right) & \text{otherwise} \end{cases}$$
ii)  
\[
\hat{k}_{(2)} = \begin{cases} 
0 & \text{if } S_i^2 + \max(0, MSB - MSE) = 0 \\
\max(0, 1 - S_i^2 / \max(0, MSB - MSE)) & \text{otherwise}
\end{cases}
\]

iii)  
\[
\hat{k}_{(3)} = \begin{cases} 
0 & \text{if } \frac{nm(\bar{Y}_{i} - \bar{Y})^2}{n-1} = 0 \\
\min\left[\max\left(0, \frac{MSB - MSE}{nm(\bar{Y}_{i} - \bar{Y})^2 / (n-1)}\right), 1\right] & \text{otherwise}
\end{cases}
\]

and

iv)  
\[
\hat{k}_{(4)} = \begin{cases} 
0 & \text{if } S_i^2 + \max(0, MSB - MSE) = 0 \\
\min\left[\max\left(0, \frac{nm(\bar{Y}_{i} - \bar{Y})^2}{n-1} - S_i^2}{S_i^2 + \max(0, MSB - MSE)}\right), 1\right] & \text{otherwise}
\end{cases}
\]

The estimators \( \hat{k}_{(2)} \) and \( \hat{k}_{(3)} \) are based on the method of moments, i.e., equating observed and expected mean squares. Since

\[
MSB = (n-1)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} (\bar{Y}_{i} - \bar{Y})^2 = \frac{1}{n} \sum_{i=1}^{n} nm(\bar{Y}_{i} - \bar{Y})^2_{n-1},
\]

and

\[
MSE = \left[ (n(m-1))^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_{i})^2 \right] = \frac{1}{n} \sum_{i=1}^{n} \left[ (m-1)^{-1} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_{i})^2 \right].
\]

the following estimators can be proposed:

a) for \( a_i^2 \),

\[
Silvina&al2005v8b_ed.doc - 07/06/05
\]
\[ \hat{a}_i^2 = \frac{\sum_{j=1}^{m} (Y_{ij} - \bar{Y}_i)^2}{m-1} = S_i^2, \]

b) for \( m \hat{b}_i^2 \),

\[ \hat{m} \hat{b}_i^2 = \max(0, \text{MSB} - \text{MSE}) \]

and

c) for \( a_i^2 + m \hat{b}_i^2 \) (considering (2.8) or (2.9), respectively)

\[ (a_i^2 + m \hat{b}_i^2)_1 = S_i^2 + \max(0, \text{MSB} - \text{MSE}). \]  

(2.10)

or

\[ (a_i^2 + m \hat{b}_i^2)_2 = \frac{nn((\bar{Y} - \bar{Y})^2)}{n-1}, \]  

(2.11)

This estimators are then combined to form \( \hat{k}_{(2)} \) and \( \hat{k}_{(3)} \).

The estimators \( \hat{k}_{(1)} \) and \( \hat{k}_{(4)} \) take into account the possibility of variance heterogeneity in the random effects and are included for comparison purposes. They consider estimating \( m \hat{b}_i^2 \) as

\[ \hat{m} \hat{b}_i^2 = \max \left(0, \frac{nn((\bar{Y} - \bar{Y})^2)}{n-1} - S_i^2 \right). \]

Finally, to obtain the empirical predictors, we also need to estimate

\[ \mu = \frac{\sum_{j=1}^{m} \frac{\bar{Y}_i}{(a_i^2 + m \hat{b}_i^2)}}{\sum_{i=1}^{n} \frac{1}{(a_i^2 + m \hat{b}_i^2)}}. \]  

From (2.10) and (2.11) we derive two estimators, namely...
\[ \hat{\mu}_1 = \frac{\sum_{i=1}^{n} \frac{Y_i}{(a_i^2 + mb_i^2)_1}}{\sum_{i=1}^{n} 1/(a_i^2 + mb_i^2)_1} \] (which is combined with \( \hat{k}_{(2)} \) or \( \hat{k}_{(4)} \) to form two predictors) and

\[ \hat{\mu}_2 = \frac{\sum_{i=1}^{n} \frac{Y_i}{(a_i^2 + mb_i^2)_2}}{\sum_{i=1}^{n} 1/(a_i^2 + mb_i^2)_2} \] (which is combined with \( \hat{k}_{(1)} \) or \( \hat{k}_{(3)} \) to form two additional predictors).

3. Details of the simulation study

We conducted a simulation study to compare the MSE of different predictors in the context of a two-stage cluster sample from a balanced finite population. The simulation study is carried out in three steps: 1) generation of the finite population, 2) selection of two stage cluster samples from the finite population, and 3) evaluation of predictors and estimated predictors for comparative purposes.

3.1. Generation of the finite populations

To encompass a broad number of situations, different compositions of the finite populations are considered. They differ with respect to: 1) the number of clusters, \( N \) and
the number of units within clusters, $M$, 2) the shape of the response variable distribution, 3) the between cluster variance, $\sigma^2$ and 4) the assumptions of equal or different within cluster variances. The presence (or not) of response error is also considered.

We define each population via a set of values based on percentiles of a hypothetical distribution and its percentiles to create a population of units and clusters. The basic distributions from which we generate the finite populations are normal, uniform, beta, or gamma. Although different distributions can be selected for units and clusters, we use the same distribution to generate the unit effects for all clusters in each population. The cluster distribution may or not agree with that for the units.

For each simulation, the population is composed of $N$ clusters with $M$ units per cluster. We represent each individual cluster parameter by $\mu_i$ and their mean by $\mu$. We fix the variance between cluster parameters, $\sigma^2$, divide the [0,1] interval into $N+1$ equally spaced intervals and obtain the percentiles corresponding to the upper limit of each interval from the appropriate probability distribution. We redefine the cluster parameters by centering them at $\mu$ and re-scaling their values so that the variance matches

$$\sigma^2 = \frac{\sum_{i=1}^{N} (\mu_i - \mu)^2}{N-1}.$$ 

Next, we generate unit effects for the $M$ units within each cluster using percentiles of a specified distribution (normal, uniform, beta, or gamma) and force these effects to average zero. The variance of the unit effects may be either set to be constant for all clusters or to vary proportionally to $\max(0, 1 - \frac{\mu_i}{\mu})$. The parameters for the cluster units
are formed by adding the unit effect to the cluster mean and are represented by $y_{iu}$. The variance of the unit parameters in cluster $s$ is $\sigma^2 = \frac{1}{M-1} \sum_{i=1}^{M} (y_{iu} - \mu_i)^2$. The common within cluster variance is equal to the average within cluster variance, i.e., $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_{iu} - \mu_i)^2$. Unit effects are re-scaled so that they have zero mean for each cluster and average variance equal to $\sigma^2$.

Using $\sigma^2$, $\sigma^2$ and $\sigma^2$ we define the cluster intra-class correlation coefficient as $\rho_s = \frac{\sigma^2}{\sigma^2 + \sigma^2}$ and the unit intra-class correlation coefficient as $\rho_i = \frac{\sigma^2}{\sigma^2 + \sigma^2}$. Note that $\rho_i = 1$ corresponds to the case with no response error.

A SAS macro (PSGEN) was developed to generate the finite populations as described in this section and is available from http://www.ime.usp.br/~jmsinger. The user may specify the parameters used in the simulation.

The characteristics of the simulated populations are summarized in Table 1. Eighty-four populations were generated in case 1, corresponding to all combinations of within cluster variances and cluster and unit intra-class correlation coefficients. In cases 2 and 3, 252=(6x6x7) populations were generated and 84=(2x6x7) populations were generated in cases 4 and 5. In total, 756 populations were evaluated.

**Insert Table 1 here**
3.2. Two-stage cluster sampling

Using a list of cluster labels, a simple random sample without replacement of \( n \) cluster labels is identified for each generated population. The identified sample clusters are combined with the population data, and from these data, a simple random sample without replacement of \( m \) units in each sampled cluster is selected. When response error is considered, it is added to the unit parameters \( y_{it} \) during the selection of the two-stage samples. We refer to this entire process as a ‘trial’.

For each generated population under cases 1, 2, and 3 (Table 1), three cluster sampling fractions \( (F = n/N = 0.2, 0.5 \text{ and } 0.8) \) and three unit sampling fractions \( (f = m/M = 0.4, 0.6 \text{ and } 0.8) \) are considered. This results in nine sampling plans for each generated population. For cases 4 and 5 (Table 1), three cluster sampling fractions \( (F = 0.2, 0.5 \text{ and } 0.8) \) and seven unit sampling fractions \( (f = 0.1, 0.2, 0.4, 0.5, 0.6, 0.8 \text{ and } 0.9) \) are considered, resulting in 21 sampling plans for each generated population. For each population and sampling plan, the number of trials is 10000. In total, 8820 different population/sampling plans were simulated.

3.3. Predictors

Once the two-stage samples are obtained, we compute the predictors and empirical predictors as detailed in Sections 2.4 and 2.5, respectively. For the ME and the SS models we assume that the known values of \( \sigma^2 \) and \( \sigma^2_e \) correspond to the between and within
cluster variances. To check the simulation results, we compare them to the theoretical mean squared error (TMSE) from Stanek and Singer (2004) for cases with equal within cluster variances and known variances.

Under the known variance assumption the MSE is denoted by SMSE and under the unknown variance assumption it is denoted by EMSE. The latter corresponds to the empirical predictors described in Section 2.5.

A SAS macro (PRED) was developed to select the two-stage samples, obtain predictors, empirical predictors and MSE. It is available from http://www.ime.usp.br/~jmsinger.

4. Simulation Results

To clarify the results, we present the equal and different within cluster variance cases separately.

For equal within cluster variances, Stanek and Singer (2004) obtain expressions for the expected MSE for ME, SS and RP model predictors when variances are known and present a small scale simulation study to compare their performance under some settings with varying unit sampling fraction \( f \) and varying cluster and unit correlation coefficients \( \rho_c \) and \( \rho_u \). We expand that simulation study by also varying cluster sampling fractions and evaluating the performance of the cluster mean (CM) as well. We include expected mean squares for the CM and for each model predictors in order to give a theoretical framework.
to put the new results (concerning unequal variances and settings where variances are estimated) in context. When within cluster variances are equal but unknown, we evaluate the loss corresponding to the use of empirical predictors and we determine under what settings each empirical predictor has the best performance.

For different within cluster variances, as optimality properties are not yet developed analytically, we consider the performance of predictors under each model when variances are known or unknown.

Due to space limitations, only some findings are presented in the Appendix; complete tables and high quality figures can be obtained from http://www.ime.usp.br/~jmsinger. The plots are organized in such a way that unit intra-class correlation coefficient $\rho_u$ increases from left to right and cluster intra-class correlation coefficient $\rho_c$ increases from top to bottom. In general, the Figures will give more detailed information including the intra-class correlation coefficients as well as the sampling fractions, while the Tables will summarize the results across all the sampling fractions.

1) Equal within cluster variances

   a) Known variances
From the expressions derived by Stanek and Singer (2004), predictors obtained under the \( RP \) model have minimum \( TMSE \), independently of the number of clusters, number of units or values of cluster or unit intra-class correlation coefficients. The expected \( TMSE \) for each predictor does not depend on the shape of the response distribution and it decreases when the cluster or unit sampling fractions increase.

For the sampling fractions considered here, Table A.1. and Figures A.1 to A.4 in the Appendix show the (maximum) percent increase in \( TMSE \) when using the \( CM \), or the \( ME \) and \( SS \) model predictors.

The predictors under the \( ME \) model are closer to those under the \( RP \) model, having a lower (maximum) relative percent increase in \( TMSE \) than the \( CM \) or the \( SS \) model predictors. As the number of clusters and the cluster sampling fraction increase, the relative percent increase (RPI) in \( TMSE \) also increases.

Despite a few departures (depending on the characteristics of the populations as well as the sampling fractions), we identify general settings for the intra-class correlation coefficients where the \( CM \), the \( ME \) and the \( SS \) model predictors have similar (RPI < 5%) as well as inferior (RPI > 50%) performance relative to \( RP \) model predictors (see Figures A.1 to A.3 as well as Table 2).

**Insert Table 2 here**

From Table 2 we observe that all models have similar performance when both intra-class correlation coefficients are high (\( \rho_c \geq 0.95 \) and \( \rho_u \geq 0.8 \)). Moreover, all predictors
behave similarly with respect to the $TMSE$ for $\rho_i \geq 0.2$ when $\rho_i = 0.99$. The performance of the $ME$ model predictor is most similar to that of the $RP$ model predictor ($RPI < 8\%)$ under a wider range of conditions (i.e. $\rho_i \geq 0.95$ or $\rho_i \leq 0.5$), followed by the $SS$ model predictor when both intra-class correlation coefficients vary (from $\rho_i \geq 0.2$ and $\rho_i \geq 0.95$ to $\rho_i \geq 0.99$ and $\rho_i \geq 0.05$ with one correlation coefficient increasing as the other decreases). Apart from the case where all predictors can be considered equivalent, the $CM$ has similar $TMSE$ to the $RP$ model predictor when $\rho_i \geq 0.8$ and $\rho_i \geq 0.8$, but only for high unit sampling fractions ($f \geq 0.6$) (not shown in Table 2).

Predictors derived under the $ME$ model have poor performance when the cluster intra-class correlation coefficient is small ($\rho_i \leq 0.2$), the unit intra-class correlation coefficient is large ($\rho_i \geq 0.95$) and the unit sampling fraction is large ($f \geq 0.8$) (see Figures A.2 and A.4). Both the $SS$ model predictor and the $CM$ present higher $TMSE$ as intra-class correlation coefficients tend (jointly) to zero, but this happens for increasing $f$ in the case of the $SS$ model predictor and for decreasing $f$ in the case of the $CM$ (see Figures A.1, A.3 and A.4).

We also use the theoretical results to check the simulation process. With 10,000 trials, the maximum relative percent difference between simulated and theoretical MSE was 2.7%.

**b) Unknown variances**
To evaluate the loss associated to the use of empirical predictors instead of those where variance components are known, we compute the RPI of the \( \text{EMSE} \) with respect to the \( \text{TMSE} \) for each predictor. [I think this would be most useful if expressed relative to the \( \text{TMSE} \) for the RP model. In this way, a smaller percent increase would indicate a better predictor. We could also summarize the proportion of time one predictor is better.] A summary of the results is presented in Table 3.

Insert Table 3 here

In this context, observe that the \( \text{EMSE} \) generally overestimates the \( \text{TMSE} \). The \( \text{CM} \) presents the lowest maximum RPI. Apart from it, the best performance is attained for empirical predictors under the \( SS \) model, with \( |\text{RPI}| < 20\% \) for about 80\% of the cases, while for the \( ME \) and \( RP \) predictors, this percentage reduces to about 55\%. The worst performance for all models is attained when unit-sampling fractions are small, especially when both the intra-class correlation coefficients decrease. An exception occurs for the \( ME \) model empirical predictors, where this poor performance is also observed for high unit sampling fractions when \( \rho_i \geq 0.8 \) and for varying values of \( \rho_s \) depending on population and cluster sampling fractions (see Figure A.5). In view of such results, we expect that:

a) the settings where all models are equivalent be maintained,

b) the settings where the \( ME \) and \( RP \) model predictors are equivalent be maintained, except for some cases where the \( ME \) model predictors would exhibit better performance than the \( RP \) model predictors,
c) the number of settings where the CM and the SS model predictors perform well increases,

d) the superiority of RP model predictors is not uniformly preserved.

We identify certain patterns (depending on $\rho_r$ and $\rho_t$) where models can be considered equivalent in terms of EMSE, although there are slight departures depending on the number and size of clusters and on the cluster or unit sampling fractions. Table 4 summarizes these results ($|\text{RPI}| < 10\%$). For the CM and the SS model predictors, the overall minimum RPI is -12.5% and -14%, respectively. In contrast, the minimum overall RPI for predictors derived under the ME model is -42%, indicating the superiority of these predictors under certain settings. The ME model predictors have better performance (RPI < -20%) than the RP model predictors generally when $\rho_r = 0.01$, $0.2 \leq \rho_t \leq 0.8$, $f$ is in the proximity of 0.5 and $F \geq 0.5$.

Insert Table 4 here

As expected, when we evaluate the loss in terms of MSE, the region where the CM and the SS model predictors are equivalent to the RP model predictors is expanded relatively to the theoretical framework, while it is reduced for the ME model predictors, sometimes with a better performance than the RP model predictors.

Particularly, considering the minimum MSE, when the number of clusters or the cluster-sampling fraction is small, the RP model predictors have the best performance in only a few situations ($f$, $\rho_r$ and $\rho_t$ small), but as the number of clusters or the cluster
sampling fraction increase, it becomes the best model for a wider range of conditions (\(f\) moderate or large and a broader range of intra-class correlations). For example, for populations with \(N = 10\), \(M = 5\) and considering \(F = 0.2\), the RP model predictors have the best performance for \(\rho_i \leq 0.2\), \(\rho_i \leq 0.05\) and any \(f\), but only for \(f = 0.4\) when \(\rho_i\) increases to 0.2. On the other hand, for \(N = 50\) and \(F = 0.8\), it is the best model for \(f \geq 0.2\), \(\rho_i \geq 0.2\) and any \(\rho_i\).

The predictors with minimum \(EMSE\) do not change when we modify the response distribution.

2) **Different within cluster variances**

   a) **Known variances**

   The predictors derived under the RP model always have the smallest \(SMSE\). Here, we also identify settings where the CM, and the ME and SS model predictors have similar (\(|RPI| < 10\%\) performance relative to RP model predictors. We summarize the findings in Table 5, from which we observe that the pattern is very similar to the case of equal within cluster variances displayed in Table 2.

   **Insert Table 5 here**
We also identify settings where the CM and the ME and SS model predictors have inferior (RPI > 50%) performance relative to RP model predictors. These settings are analogous to the case of equal within cluster variances.

b) Unknown variances

The CM has a similar performance (|RPI| < 10%) to RP model predictors when both intra-class correlations are high (\( \rho_s \geq 0.95 \) and \( \rho_t \geq 0.8 \), extending to \( \rho_s \geq 0.05 \) when \( \rho_s = 0.99 \)). Although in some settings, for populations with \( N = 10 \) and cluster sampling fraction \( F = 0.2 \), the best predictor is the CM, it presents a minimum RPI of -12.6%, so that its performance can be considered equivalent to that of the RP model predictors.

With respect to the ME and SS model predictors, none of the proposed estimated shrinkage constants reproduce the results of Table 5. The ME and SS model predictors using \( \hat{k}_4 \) are generally bad as well as the ME model predictors with \( \hat{k}_1 \). Although the predictors derived under ME and SS models using \( \hat{k}_3 \) sometimes have better performance than that of the RP model predictors (in particular for populations with \( N = 10 \) and cluster sampling fraction \( F = 0.2 \)), the results seem to be unstable. When both intra-class correlation coefficients are large, the performance of such predictors become very poor, with EMSE much greater than that of the other predictors (see Figure A.7 and A.8). For populations with \( M = 5 \), the ME model predictors using \( \hat{k}_3 \) present a minimum RPI of -20%, -31% and -38% for cluster sampling fractions of 0.2, 0.5 and 0.8, respectively, while
for populations with $M = 20$, these predictors present a minimum RPI of -17%, -24% and -29% for cluster sampling fractions $F = 0.2, 0.5$ and $0.8$, respectively. The SS model predictors using $\hat{k}_i$ present an overall minimum RPI of -16%. In all situations, the second best predictors are always those derived under the RP model.

Particularly, the RP model-based predictors have better performance under a wider range of conditions as either the number of clusters or the sampling fraction of clusters increase.

5. Conclusions

When variances are known, the RP model produces the best predictors in terms of MSE no matter if the within cluster variances are equal or different. For populations with identical within cluster variances, the shape of the response variable distribution has almost no effect on the values of the MSE. Although here we only consider some distributions for continuous variables, another study is currently being developed employing distributions for discrete variables.

When variances are unknown, the performance of the predictors obtained under the RP model improves as the number and size of clusters increase and as the cluster sampling fraction also increases, becoming the best over a wider range of conditions.

As $\rho_i$ and $\rho_j$ gradually increase, predictors derived under the ME, the SS model or the CM, respectively, can be considered equivalent ($|\text{RPI}| < 10\%$) to those obtained.
under the \( RP \) model. When both intra-class correlation coefficients are very high
\((\rho_i \geq 0.95 \text{ and } \rho_j \geq 0.8)\), the performance of predictors derived under the three models and
the \( CM \) can be considered equivalent (\(|RPI| < 10\%\)). These results are valid when a) the
within cluster variances are equal or b) they are different but known.

When within cluster variances are different and unknown, there is still a need of
more study. From the simulation results, it seems that, although the \( RP \) model predictors do
not perform as well as the \( ME \) or \( SS \) model predictors defined with shrinkage constant \( \hat{k} \),
in some situations, the latter exhibit very poor performance when the sampling fractions
and the intra-class correlation coefficients increase, while the former shows a more stable
performance, being the best or the second best.

In summary, we may recommend the \( RP \) model predictors, except in cases where all
the following conditions are satisfied: variances are unknown, the cluster-sampling fraction
is small (\( F = 0.2 \)), the cluster intra-class correlation coefficient is very small (\( \rho_i = 0.01 \))
and the unit intra-class correlation is moderate to large (\( 0.5 \leq \rho_j \leq 0.95 \)).

Finally, as in any simulation study, the results obtained are empirical and additional
research may be necessary to obtain more definite conclusions.

Acknowledgements

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

References


APPENDIX – PERFORMANCE OF PREDICTORS UNDER \textit{CM, ME, SS AND RP} MODELS

1) Equal and known within cluster variances

Insert Table A.1. here

Insert Fig. A.1. here

Insert Fig. A.2. here

Insert Fig. A.3. here

Insert Fig. A.4. here

2) Equal and unknown within cluster variances

Insert Fig. A.5. here

Insert Fig. A.6. here
3) Different and unknown within cluster variances

Insert Fig. A.7. here

Insert Fig. A.8. here
Table captions

Table 1. Characteristics of the simulated populations

Table 2. Settings (depending on the intra-class correlation coefficients) where the CM, the ME and the SS model predictors have equivalent TMSE (RPI < 5%) to the RP model predictors

Table 3. Relative percent increase (RPI) between EMSE and TMSE for each predictor

[Titles for Table 4 and 5 are the same]

Table 4. Settings where the CM, the ME and the SS model predictors have similar (|RPI| < 10%) performance relative to the RP model predictors

Table 5. Settings where the CM, the ME and the SS model predictors have similar (|RPI| < 10%) performance relative to RP model predictors

Table A.1. Maximum relative percent increase (RPI) in expected TMSE for the CM, the ME and the SS model predictors relative to the RP model predictors for each population
Figure captions

Figure A.1. Relative percent increase in expected $TMSE$ for the $CM$ relative to the $RP$ model predictors for the population with $N=10$, $M=20$ and equal and known within cluster variances. Axis named $f$ takes values 0.1, 0.2, 0.4, 0.5, 0.6, 0.8 and 0.9. Axis named $F$ takes values 0.2, 0.5 and 0.8 and maximum value for axis named % is 700.

Figure A.2. Relative percent increase in expected $TMSE$ for the $ME$ model predictors relative to the $RP$ model predictors for the population with $N=10$, $M=20$ and equal and known within cluster variances. Axis named $f$ takes values 0.1, 0.2, 0.4, 0.5, 0.6, 0.8 and 0.9. Axis named $F$ takes values 0.2, 0.5 and 0.8 and maximum value for axis named % is 70.

Figure A.3. Relative percent increase in expected $TMSE$ for the $SS$ model predictors relative to the $RP$ model predictors for the population with $N=10$, $M=20$ and equal and known within cluster variances. Axis named $f$ takes values 0.1, 0.2, 0.4, 0.5, 0.6, 0.8 and 0.9. Axis named $F$ takes values 0.2, 0.5 and 0.8 and maximum value for axis named % is 450.

Figure A.4. Relative percent increase (RPI) in expected $TMSE$ for the $CM$, the $ME$ and the $SS$ model predictors relative to the $RP$ model predictors for the population with $N=10$, $M=20$ and equal and known within cluster variances. Axis named $f$ takes values 0.1, 0.2, 0.4, 0.5, 0.6, 0.8 and 0.9. Axis named $F$ takes values 0.2, 0.5 and 0.8 and maximum value for axis named % is 450.
\( M = 5 \), equal and known within cluster variances and cluster sampling fraction \( F = 0.2 \). Darker gray indicates increasing values of RPI according to the scale shown on the left.

Figure A.5. Relative percent increase in \( EMSE \) relative to \( TMSE \) for predictors derived under each model for the population with \( N = 10 \), \( M = 20 \), equal within cluster variances and cluster sampling fraction \( F = 0.5 \).

Figure A.6. Relative percent increase in \( EMSE \) relative to \( EMSE \) for \( RP \) model predictors for the population with \( N = 10 \), \( M = 5 \), equal and unknown within cluster variances and cluster sampling fraction \( F = 0.5 \).

Figure A.7. Relative percent increase in \( EMSE \) for the \( CM \) and the \( ME \) model predictors relative to \( EMSE \) for \( RP \) model predictors for the population with \( N = 10 \), \( M = 5 \), different and unknown within cluster variances and cluster sampling fraction \( F = 0.5 \).

Figure A.8. Relative percent increase in \( EMSE \) for the \( CM \) and the \( SS \) model predictors relative to \( EMSE \) for \( RP \) model predictors for the population with \( N = 10 \), \( M = 5 \), different and unknown within cluster variances and cluster sampling fraction \( F = 0.5 \).

---

4 Note: Due to the great difference in magnitude, each plot has different scales
5 Note: Due to the great difference in magnitude, each plot has different scales