A SPREE SMALL AREA PROCEDURE FOR ESTIMATING POPULATION COUNTS

Emily Berg[1] and Wayne A. Fuller[2]

ABSTRACT

The Canadian Labour Force Survey uses a hierarchical system to classify employment into occupations. Reliable estimates of occupational totals at the province level are desired. A small area procedure is developed under the assumption that the direct estimators of the margins of the two-way table (province and occupations) have acceptable reliabilities. The procedure obtains more stable estimates of the counts in the two-way table and preserves the estimates of the marginal totals. Auxiliary information from the previous Census is incorporated through the SPREE estimates of the cell totals. Estimators of the random model are used to combine the SPREE estimates with the direct estimates.

KEY WORDS: Benchmarking, Small area estimation, SPREE.

RÉSUMÉ

L'Enquête sur la population active du Canada se sert d'un système hiérarchique afin de classifier l'emploi en occupations. Nous développons une méthode SPREE pour les petites régions en supposant que les marges du tableau bidimensionnel, province et occupations, aient des fiabilités acceptables. Cette méthode nous permet d'avoir des estimations plus stables pour les comptes à l'intérieur de ce tableau bidimensionnel tout en respectant les totaux marginaux. De l'information auxiliaire provenant les données d'un recensement précédent est incorporée aux estimateurs des totaux SPREE. On se sert d'estimateurs basés sur un modèle aléatoire afin de combiner les estimations des proportions provenant des totaux découlant des estimateurs SPREE avec les totaux dérivés à partir d'estimations directes.

MOTS CLÉS : Estimations sur petits domaines; l'étalonnage; SPREE.

1. INTRODUCTION

1.1 Description of the Problem

The Canadian Labour Force Survey (LFS) produces monthly estimates of employment characteristics and standard labour market indicators. The LFS collects occupational data using a hierarchical classification system in which categories labelled with three digit codes are nested in categories labelled with two digit codes. As an example, A1 is the two digit code that denotes the category for specialist managers. The four three digit codes, A11-A14, partition specialist managers into more specific occupations.

Because occupations are unplanned domains in the LFS sample design, realized sample sizes in occupations are random. Small realized sample sizes cause the direct estimators of three digit totals and proportions in small provinces to have unacceptably large coefficients of variation. In contrast, the direct estimators of the province two digit totals and the national three digit totals (within two digit codes) are judged to have adequate design properties in terms of both bias and precision. (Hidiroglou and Patak, 2009)

The Canadian Census of Population, conducted every five years, publishes occupational counts through the three digit level of detail for each province. The proportions of two digit codes in each three digit code calculated with the Census data by province are highly correlated with the corresponding proportions calculated with the direct LFS estimators. Although differences between the data collection protocols used in the LFS and the Census may lead to some differences in the resulting occupational data, the Census provides a good source of auxiliary information for estimating three digit occupational totals at the province level with the LFS data. (Hidiroglou and Patak, 2009)

[1] Emily Berg, Center for Survey Statistics and Methodology, Iowa State University, Ames, IA, US, 50014, emilyb@iastate.edu
[2] Wayne A. Fuller, Center for Survey Statistics and Methodology, Iowa State University, Ames, IA, US, 50014, waf@iastate.edu

Our objective is to obtain estimators of the cell totals in the two-way table defined by the cross-classification of provinces and three digit codes in a single two digit code with better precisions than the direct estimators. Because the direct estimators of the margins of the two-way table have acceptable design coefficients of variation, we desire predictors of the cell totals that preserve the direct estimators of the margins. We will also make use of the Census two-way table to the extent that the Census data provide useful information about the current employment totals.

1.2 Organization of the Paper

Section 2 briefly reviews model based approaches to small area estimation that are related to the procedure that we suggest. In section 3, we specify a model for the direct LFS estimators. The model consists of a nonlinear expectation function and random errors. In section 4, we describe a model based predictor and define an estimator of the mean squared error. We evaluate the procedure through simulation in section 5. A summary is in section 6.

2. LITERATURE REVIEW

The challenges associated with obtaining reliable estimators of the employment totals at the three digit level by province are characteristic of the difficulties encountered in "small area estimation," a term that describes estimation for domains (or small areas) in which realized sample sizes are too small to produce stable direct estimators. A widely adopted solution uses models to combine direct estimators with synthetic estimators in a model-optimal fashion.

2.1 Model Based Small Area Estimation

Models for small area estimation typically have a hierarchical structure in which a "sampling model" describes the conditional distributions of the direct estimators given the true values, and a "linking model" (Liu, et al., 2007) connects the true values to a set of auxiliary variables. Fay and Herriot (1979) use the linear mixed model with normally distributed random effects and an assumption of a known sampling variance. Numerous applications extend the Fay-Herriot (1979) procedure to models with nonlinear expectation functions or non-normal error distributions. Rao (2003) and Jiang and Lahiri (2006) review methods for small area estimation.

2.2 SPREE and Generalizations

We desire estimators for cell totals and column proportions of a two-way table that preserve the direct estimators of the marginal totals. Purcell and Kish (1980) propose Structure Preserving Estimation (SPREE) as a way to update the contingency table from a previous census to direct estimators of the marginal totals obtained from a current survey. Noble et al. (2002) demonstrate that SPREE is maximum likelihood estimation under a generalized linear model in which the interactions in the time point of interest are set equal to the interactions in the Census. If the loglinear expectation function underlying SPREE fails to hold, the SPREE estimators can be severely biased (Griffiths, 1996; Zhang and Chambers, 2004). Griffiths (1996) develops a composite estimator that is a weighted combination of the direct estimator and the SPREE estimator. Zhang and Chambers (2004) first define a class of loglinear models called generalized linear structural models in which the interactions are proportional to the Census interactions. They then extend the generalized linear structural model to a mixed model with normally distributed random effects. We extend the SPREE model to a different nonlinear mixed model.

2.2 Semi-Parametric Models for Small Area Estimation

The model in section 3 specifies first and second moments of the random errors. Other examples of model based approaches to small area estimation that avoid full distributional assumptions include Longford (1999), Ghosh and Lahiri (1987), and Raghunathan (1993). Longford defines a predictor of a vector of proportions that estimates the minimum mean squared error linear combination of a vector of area-specific estimators and a national average. Ghosh and Lahiri (1987) define related predictors under an assumption of posterior linearity. Raghunathan (1993) derives predictors under quasi-posterior predictive distributions based on quasi-likelihoods.

3. MODEL

We specify a model for the relationship between the Census data, the true proportions and totals, and the direct estimators. Let $\hat{p}_{ik}$ be the direct estimator of the ratio of the total in province k employed in three digit code i to the two digit total for province k, where $i = 1, \ldots, I$, and $k = 1, \ldots, K$. Assume that the direct estimator of the proportion satisfies

$$\hat{p}_{ik} = p_{T,ik} + u_{ik} + e_{ik}, \tag{1}$$

where u_{ik} is a mean zero small area effect, e_{ik} is a mean zero sampling error, and

$$p_{T,ik} = h_{ik}(\boldsymbol{N}, \boldsymbol{T}_{I.}, \boldsymbol{T}_{.K}) \tag{2}$$

is a function of the vector of Census totals, $\boldsymbol{N}$, the vector of the expected values of the direct estimators of the national three digit totals, $\boldsymbol{T}_{I.} = (T_{1.}, \ldots, T_{I.})'$, and the vector of the expected values of the direct estimators of the province two digit totals, $\boldsymbol{T}_{.K} = (T_{.1}, \ldots, T_{.K})'$. Assume that u_{ik} and e_{jt} are uncorrelated for all i, k, j, and t. Because we treat the Census totals as fixed, $p_{T,ik}$ is a fixed parameter. The true cell proportion to be predicted is

$$p_{ik} = p_{T,ik} + u_{ik}, \tag{3}$$

and the corresponding true cell total is $M_{ik} = p_{ik} T_{.k}$. We discuss the expectation function in (2) and the moments of the random errors in the next subsections.

3.1 Expectation Function

We assume that the function $h_{ik}(\boldsymbol{N}, \boldsymbol{T}_{I.}, \boldsymbol{T}_{.K})$ in (2) transforms the vector of Census totals and the expected values of the direct estimators of the marginal totals in a way that preserves the odds ratios in the two-way table of Census totals and the margins specified by $\boldsymbol{T}_{I.}$ and $\boldsymbol{T}_{.K}$. One way to characterize the function in (2) is through the operation that rakes the two-way table of Census totals to the expected values of the direct estimators of the national three digit and province two digit totals. Let T_{ik} be the entry in row i and column k of the two-way table obtained on convergence of the iterative raking procedure. The row and column sums of the two-way table with T_{ik} in row i and column k are equal to $\{T_{i.} : i = 1, \ldots, I\}$ and $\{T_{.k} : k = 1, \ldots, K\}$, respectively. The expected value of the direct estimator of the proportion, defined in (2), is $p_{T,ik} = h_{ik}(\boldsymbol{N}, \boldsymbol{T}_{I.}, \boldsymbol{T}_{.K}) = T_{.k}^{-1} T_{ik}$.

An equivalent way to describe the function in (2) is through a loglinear model. Consider a saturated loglinear model specifying the Census totals to have independent Poisson distributions with means $\{\mu_{ik}^{cen} : i = 1, \ldots, I; k = 1, \ldots, K\}$ satisfying, $\log(\mu_{ik}^{cen}) = \alpha_{o,i}^{cen} + \beta_{o,k}^{cen} + (\alpha\beta)_{o,ik}^{cen}$. For uniqueness, set $\alpha_{o,1}^{cen} = (\alpha\beta)_{o,1k}^{cen} = (\alpha\beta)_{o,i1}^{cen} = 0$ for all i and k. Let $\{(\alpha\beta)_{ik}^{cen} : i = 1, \ldots, I; k = 1, \ldots, K\}$ be the maximum likelihood estimates of the interactions, where $(\alpha\beta)_{1k}^{cen} = (\alpha\beta)_{i1}^{cen} = 0$ for all i and k. Then, an alternative expression for T_{ik}, defined above, is

$$T_{ik} = \exp(\alpha_{o,i} + \beta_{o,k} + \alpha\beta_{ik}^{cen}), \tag{4}$$

where $\alpha_{o,1} = 0$, and $\alpha_{o,i}$ and $\beta_{o,k}$ satisfy $T_{i.} = \sum_{k=1}^{K} T_{ik}$, for $i = 1, \ldots, I$ and $T_{.k} = \sum_{i=1}^{I} T_{ik}$, for $k = 1, \ldots, K$. As above, $p_{T,ik} = h_{ik}(\boldsymbol{N}, \boldsymbol{T}_{I.}, \boldsymbol{T}_{.K}) = T_{.k}^{-1} T_{ik}$. By construction, $\sum_{i=1}^{I} p_{T,ik} = 1$.

The model (4) for T_{ik} is the expectation function in a loglinear model underlying SPREE (Noble, et al. 2002). The function $h_{ik}(\boldsymbol{N}, \boldsymbol{T}_{I.}, \boldsymbol{T}_{.K})$ retains the odds ratios, or equivalently the interactions, from the Census. Zhang and Chambers introduce a more general model, the generalized linear structural model (GLSM), which permits the interactions in the two-way table of expected values for the time point of interest to be proportional to (rather than equal to) the Census interactions. Because the procedure used in the LFS is based on the SPREE model for the expected values, we restrict our attention to the SPREE model, a special case of the GLSM.

3.2 Small Area Effects and Sampling Errors

Let the vector of small area effects for a province be denoted by $\boldsymbol{u}_k = (u_{1k}, \ldots, u_{Ik})'$, and assume that the population covariance matrix of $\boldsymbol{u}_k$ is

$$\mathrm{V}\{\boldsymbol{u}_k\} = \boldsymbol{\Gamma}_{uu,k} \sigma_u^2 = [\mathrm{diag}(\boldsymbol{p}_{T,k}) - \boldsymbol{p}_{T,k} \boldsymbol{p}_{T,k}'] \sigma_u^2, \tag{5}$$

where $\boldsymbol{p}_{T,k} = (p_{T,1k}, \ldots, p_{T,Ik})'$, and $\mathrm{diag}(\boldsymbol{p}_{T,k})$ is the diagonal matrix with the vector $\boldsymbol{p}_{T,k}$ on the diagonal. The variance of $\boldsymbol{u}_k$ is proportional to the multinomial covariance matrix. Because the columns of the covariance matrix in (5) sum to zero, the sum of the small area effects in a province is zero. Also, assume that the small area effects in different provinces are uncorrelated so that the variance of the vector of small area effects, $\boldsymbol{u} = (\boldsymbol{u}_1', \ldots, \boldsymbol{u}_K')'$, is block-diagonal with $\boldsymbol{\Gamma}_{uu,k} \sigma_u^2$ as the blocks.

Assume that $E[e_{ik}|u_{ik}] = 0$ so that the direct estimators are conditionally unbiased for the true proportions defined in (3). Let $\boldsymbol{\Sigma}_{ee,k}$ be the variance of $\boldsymbol{e}_k$, where $\boldsymbol{e}_k$ is the vector of design errors for province k. In the LFS, provinces are unions of design strata. Therefore, sampling is independent across provinces, and the variance of $\boldsymbol{e} = (\boldsymbol{e}_1', \dots, \boldsymbol{e}_K')'$ is block-diagonal.

3.3 Totals

The model for the direct estimator of the proportion has implications for the first and second moments of the direct estimator of the cell total. Let $\widehat{M}_{.k}$ be the direct estimator of the two digit total in province k. Assume that $\widehat{M}_{.k}$ is independent of $\boldsymbol{u}$ and that $E[u_{ik}e_{ik}|\,\dot{M}_{.k}] = 0$ for all k. Let $\dot{M}_{ik}$ be the direct estimator of the total in three digit code i and province k. Because the direct estimators of the proportions and totals satisfy $\widehat{M}_{ik} = \hat{p}_{ik}\widehat{M}_{.k}$, it follows that

$$\widehat{M}_{ik} = p_{T,ik}T_{.k} + u_{ik}\widehat{M}_{.k} + f_{ik}, \tag{6}$$

where $f_{ik} = e_{ik}\dot{M}_{.k} + p_{T,ik}(\dot{M}_{.k} - T_{.k})$, and $T_{.k}$, defined previously, is the expected value of $\widehat{M}_{.k}$. Then, f_{ik} is the sampling error in the total and is uncorrelated with $u_{ik}\widehat{M}_{.k}$. Let $\boldsymbol{d} = (\boldsymbol{e}', \boldsymbol{f}')'$ be the vector of sampling errors, where $\boldsymbol{f}$ is the $IKX1$ vector with elements $\{f_{ik}: i = 1, \dots, I;\ k = 1, \dots, K\}$ listed in the order with provinces grouped together. Using Taylor series, one can obtain an approximation for the variance of $\boldsymbol{d}$ as a function of $\{\boldsymbol{\Sigma}_{ee,k}: k = 1, \dots, K\}$, the variances of the direct estimators of the province two digit totals, and the expected values. Let $\boldsymbol{\Sigma}_{dd}$ be the approximate variance of $\boldsymbol{d}$. Let $\boldsymbol{b} = (\boldsymbol{u}', \boldsymbol{v}')'$, where $\boldsymbol{v} = (\boldsymbol{u}_1'\dot{M}_{.1}, \dots, \boldsymbol{u}_K'\dot{M}_{.K})'$. Using the assumption that $\dot{M}_{.k}$ and u_{ik} are uncorrelated, one can express the variance of $\boldsymbol{b}$ as $\sigma_u^2\boldsymbol{B}_{uu}$, where $\boldsymbol{B}_{uu}$ is a function of the expected values and the variances of the direct estimators of the province two digit totals.

4. PROCEDURE

The suggested procedure for predicting the true totals and proportions consists of several steps. First, we estimate the underlying expected value using SPREE. Second, we obtain an estimator of σ_u^2 using approximations for expected mean squares. Third, we calculate an initial predictor of the true proportion as a weighted combination of the direct estimator and the SPREE estimator, with weights determined by the estimators of the variances. Finally, we benchmark the initial predictors to the direct estimators of the marginal totals. We also define estimators of the mean squared errors of the predictors of the proportions.

4.1 Estimator of Marginal Expected Value

We estimate the expected values, T_{ik} and $p_{T,ik}$, using SPREE. The SPREE procedure evaluates the function defined in (2) at the direct estimators of the marginal totals instead of at $\boldsymbol{T}_I$ and $\boldsymbol{T}_K$. Therefore, the SPREE estimator of the proportion is

$$\hat{p}_{T,ik} = h_{ik}(N, \widehat{\boldsymbol{M}}_I, \widehat{\boldsymbol{M}}_K), \tag{7}$$

where $\widehat{\boldsymbol{M}}_I$ and $\widehat{\boldsymbol{M}}_K$ are the vectors containing the direct estimators of the national three digit totals and province two digit totals, respectively. The corresponding total is $\hat{T}_{ik} = \hat{p}_{T,ik}\widehat{M}_{.k}$, where $\widehat{M}_{.k}$, defined previously, is the direct estimator of the two digit total for province k. The SPREE estimators preserve the direct estimators of the marginal totals; $\sum_{k=1}^{K}\hat{T}_{ik} = \widehat{M}_{i.}$, and $\sum_{i=1}^{I}\hat{T}_{ik} = \widehat{M}_{.k}$, where $\widehat{M}_{i.} = \sum_{k=1}^{K}\widehat{M}_{ik}$ is the direct estimator of the national total in three digit code i.

Two ways to implement SPREE are iterative proportional fitting and Newton-Raphson. Purcell and Kish (1980) use iterative proportional fitting, a procedure that rakes the Census table to the direct estimators of the marginal totals. Use of Newton-Raphson as an alternative to iterative proportional fitting follows naturally from the characterization of SPREE as maximum likelihood estimation under a loglinear model (Noble et al., 2002; Zhang and Chambers, 2004).

Expressing the SPREE estimators as maximum likelihood estimators is also useful for obtaining a linear approximation for the SPREE estimators. If the direct estimators of the cell totals are independent Poisson random variables with means $\{T_{ik}: i = 1, \dots, I;\ k = 1, \dots, K\}$ defined in (4), then $\hat{T}_{ik}$ is the maximum likelihood estimator of T_{ik} (Noble et al., 2002; Zhang and Chambers, 2004). Letting $\boldsymbol{\lambda} = (\alpha_2, \dots, \alpha_I, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)'$ and $\alpha_1 = 0$, the score function for the generalized linear model underlying SPREE is $\boldsymbol{s}(\boldsymbol{\lambda}) = \boldsymbol{X}'(\dot{\boldsymbol{M}} - \boldsymbol{T}(\boldsymbol{\lambda}))$, where $\dot{\boldsymbol{M}}$ is the $IKX1$ vector of direct estimators of totals listed in the order with provinces grouped together, and $\boldsymbol{T}(\boldsymbol{\lambda})$ is the corresponding vector with elements

$\{T_{ik}(\boldsymbol{\lambda}) = \exp(\alpha_i + \beta_k + \alpha\beta_{ik}^{cen}): i = 1,\ldots,I; k = \mathbf{1},\ldots,K\}$. The matrix $\boldsymbol{X}$ is the IK x $(I + K - 1)$ matrix with $\boldsymbol{x}'_{ik} = (I[i = 2], \ldots, I[i = I], I[k = 1], \ldots, I[k = K])$ in row $m(k-1) + i$ (the row corresponding to three digit code i and province k), and $I[\,]$ is the indicator function, which equals 1 if the condition inside the brackets is true and equals zero otherwise. Let $\hat{\boldsymbol{\lambda}}$ satisfy $\boldsymbol{s}(\hat{\boldsymbol{\lambda}}) = \mathbf{0}$. Then, $\hat{T}_{ik} = T_{ik}(\hat{\boldsymbol{\lambda}})$, and the corresponding proportion is $\hat{p}_{T,ik} = \hat{M}_k^{-1} T_{ik}(\hat{\boldsymbol{\lambda}})$. Let $\boldsymbol{p}_T = (\boldsymbol{p}'_{T,1}, \ldots, \boldsymbol{p}'_{T,K})'$, and let $\hat{\boldsymbol{p}}_T$ be corresponding vector with elements $\{\hat{p}_{T,ik}: i = 1,\ldots,I;\ k = 1,\ldots,K\}$. Let $\boldsymbol{D}_{\boldsymbol{p}}(\boldsymbol{\lambda}_o)$ be the matrix of derivatives of $\hat{\boldsymbol{p}}_T$ with respect to $\hat{\boldsymbol{\lambda}}$ evaluated at $\boldsymbol{\lambda}_o$, where $\boldsymbol{\lambda}_o = (\alpha_{o,2}, \ldots, \alpha_{o,I}, \beta_{o,1}, \ldots, \beta_{o,K})'$ is the vector of true parameters. Then, by a Taylor expansion of the score function, $\boldsymbol{s}(\hat{\boldsymbol{\lambda}})$, around the true parameter $\boldsymbol{\lambda}_o$,

$$\hat{\boldsymbol{p}}_T - \boldsymbol{p}_T \approx \boldsymbol{W}'_T(\hat{\boldsymbol{M}} - \boldsymbol{T}), \tag{8}$$

where $\boldsymbol{W}'_T = \boldsymbol{D}_{\boldsymbol{p}}(\boldsymbol{\lambda}_o)(\boldsymbol{X}'\mathrm{diag}(\boldsymbol{T})\boldsymbol{X})^{-1}\boldsymbol{X}'$, $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{\lambda}_o)$, and $\mathrm{diag}(\boldsymbol{T})$ is the diagonal matrix with the vector $\boldsymbol{T}$ on the diagonal. Expression (8) gives an approximation for the SPREE estimators as linear functions of the direct estimators of the marginal totals. We use the linear approximation to define an estimator of σ_u^2 and estimators of the mean squared errors of the predictors.

4.2 Estimator of Small Area Effects Variance

By the linear approximation (8) and the model (1) for the direct estimators,

$$\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T \approx \boldsymbol{W}'(\boldsymbol{b} + \boldsymbol{d}), \tag{9}$$

where $\hat{\boldsymbol{p}}$ is the vector of direct estimators of proportions listed in the order with provinces grouped together, $\boldsymbol{W}' = (\boldsymbol{I}_{IK}, -\boldsymbol{W}'_T)$, and $\boldsymbol{I}_{IK}$ is the IK X IK identity matrix. Let $(\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T)^2$ be the vector containing the squares of the elements of $\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T$. By (9),

$$E\left[(\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T)^2\right] \approx \boldsymbol{\sigma}_a + \boldsymbol{\sigma}_b\sigma_u^2, \tag{10}$$

where $\boldsymbol{\sigma}_a$ and $\boldsymbol{\sigma}_b$ are the vectors containing the diagonal elements of the matrices $\boldsymbol{\Sigma}_a = \boldsymbol{W}'\boldsymbol{\Sigma}_{dd}\boldsymbol{W}$ and $\boldsymbol{\Sigma}_b = \boldsymbol{W}'\boldsymbol{B}_{uu}\boldsymbol{W}$, respectively. Under normality,

$$V\{(\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T)^2\} \approx 2(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b\sigma_u^2)^2, \tag{11}$$

where the elements of $(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b\sigma_u^2)^2$ are the squares of the elements of the matrix $(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b\sigma_u^2)$. The matrices $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_b$ are functions of the sampling variances and the expected values of the direct estimators. Let $\hat{\boldsymbol{\Sigma}}_a$ and $\hat{\boldsymbol{\Sigma}}_b$ be the estimators obtained by replacing the unknown sampling variances by LFS estimators of the sampling variances and by replacing the expected values by the SPREE estimators. Let $\hat{\boldsymbol{\sigma}}_a$ and $\hat{\boldsymbol{\sigma}}_b$ be the vectors of diagonal elements of $\hat{\boldsymbol{\Sigma}}_a$ and $\hat{\boldsymbol{\Sigma}}_b$, respectively.

An EGLS estimator of σ_u^2 under the linear model defined by (10) and (11) requires an initial value for σ_u^2. We obtain an initial estimator of σ_u^2 from an estimator of the lower bound of the variance of an EGLS estimator. If $\sigma_u^2 = 0$, then under an assumption of normally distributed errors, an estimator of the variance of an EGLS estimator of σ_u^2 is

$$\hat{V}_0 = [\hat{\boldsymbol{\sigma}}'_b(2(\hat{\boldsymbol{\Sigma}}_a)^2)^{-1}\hat{\boldsymbol{\sigma}}_b]^{-1}, \tag{12}$$

where the elements of $(\hat{\boldsymbol{\Sigma}}_a)^2$ are the squares of the elements of $\hat{\boldsymbol{\Sigma}}_a$. Under normality, the standard error of an estimator of a variance is proportional to the variance. In analogy with the method of Wang and Fuller (2003), we use $\xi = 0.5[\hat{V}_0]^{0.5}$ as the initial estimator of σ_u^2. Setting $\xi = \sigma_u^2$ in the variance expression (11), an EGLS estimator of σ_u^2 is

$$\tilde{\sigma}_u^2 = \frac{\hat{\boldsymbol{\sigma}}'_b[2(\hat{\boldsymbol{\Sigma}}_a + \hat{\boldsymbol{\Sigma}}_b\sigma_u^2)^2]^{-1}((\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T)^2 - \hat{\boldsymbol{\sigma}}_a)}{\hat{\boldsymbol{\sigma}}'_b[2(\hat{\boldsymbol{\Sigma}}_a + \hat{\boldsymbol{\Sigma}}_b\sigma_u^2)^2]^{-1}\hat{\boldsymbol{\sigma}}_b}. \tag{13}$$

We use $\hat{\sigma}_u^2 = \max(\xi, \tilde{\sigma}_u^2)$ as the final estimator of σ_u^2.

If multiple estimators of σ_u^2 are available, one can combine the estimators of σ_u^2 to obtain a more efficient estimator. For example, the LFS provides estimators of σ_u^2 from several time points, $t = 1, \ldots, d$. Let $\hat{V}_o^{(t)}$ denote the estimator (12) from the t^{th} time point, and let $\tilde{\sigma}_u^{2(t)}$ be the corresponding EGLS estimator (13), which may be negative. Then, a smoothed estimator of σ_u^2 is $\hat{\sigma}_{u(d)}^2 = \max(\bar{V}_o, \bar{\sigma}_u^2)$, where $\bar{V}_o = (2d)^{-1}[\sum_{t=1}^{d}\hat{V}_o^{(t)}]^{0.5}$, and $\bar{\sigma}_u^2 = d^{-1}\sum_{t=1}^{d}\tilde{\sigma}_u^{2(t)}$. Averaging multiple estimators of σ_u^2 has the potential to reduce the variance of the estimator of σ_u^2 and subsequently improve the precisions of the predictors.

4.3 Predictors of True Proportions and Totals

We desire predictors of the true proportions that have small mean squared errors and also preserve the direct estimators of the marginal totals. We define an initial predictor that estimates the minimum mean squared error convex combination of the direct estimator and the fixed parameter, $p_{T,ik}$. We benchmark the initial predictors to the direct estimators of the marginal totals.

A vector of predictors for province k that minimizes the mean squared error in the class of predictors that are linear in the vector of direct estimators of proportions is

$$\boldsymbol{p}_{blp,k} = \boldsymbol{p}_{T,k} + \boldsymbol{\Gamma}_{uu,k}\sigma_u^2(\boldsymbol{\Gamma}_{uu,k}\sigma_u^2 + \boldsymbol{\Sigma}_{ee,k})^{-}(\hat{\boldsymbol{p}}_k - \boldsymbol{p}_{T,k}), \quad (14)$$

where " $-$ " denotes a generalized inverse. Because the elements of $\boldsymbol{p}_{blp,k}$ can be larger than 1 or negative, we consider predictors that are convex combinations of the form, $g_{ik}\hat{p}_{ik} + (1 - g_{ik})p_{T,ik}$. The predictor that minimizes the mean squared error, $E[(g_{ik}\hat{p}_{ik} + (1 - g_{ik})p_{T,ik} - p_{ik})^2]$, is

$$\tilde{p}_{ik}(p_{T,ik}, \sigma_u^2) = p_{T,ik} + g_{ik}(\hat{p}_{ik} - p_{T,ik}), \quad (15)$$

where $g_{ik} = p_{T,ik}(1 - p_{T,ik})\sigma_u^2(\sigma_u^2 p_{T,ik}(1 - p_{T,ik}) + \sigma_{e,ik}^2)^{-1}$, and $\sigma_{e,ik}^2$ is the i^{th} diagonal element of $\boldsymbol{\Sigma}_{ee,k}$. Isaki et al. (2000) compare univariate predictors calculated with the diagonal elements of an estimated covariance matrix with the empirical best linear unbiased predictor. In their simulation study, the univariate predictors sometimes have mean squared errors smaller than the mean squared errors of predictors constructed with the full covariance matrix. If $\boldsymbol{\Sigma}_{ee,k} = \sigma_{a,k}^2\boldsymbol{\Gamma}_{uu,k}$ for some constant $\sigma_{a,k}^2$, then the univariate predictor, defined in (15), is equal to the i^{th} element of $\boldsymbol{p}_{blp,k}$, and the weight g_{ik} is constant for all categories in a province; $g_{ik} = g_k = \sigma_u^2(\sigma_u^2 + \sigma_{a,k}^2)^{-1}$.

The predictor in (15) depends on unknown parameters. To evaluate the predictor, we use the EGLS estimator of σ_u^2, the SPREE estimator of $p_{T,ik}$, and the LFS estimator of the sampling variance. We compute an initial predictor for cell i, k as

$$\hat{p}_{pred,ik} = \hat{p}_{T,ik} + \hat{g}_{ik}(\hat{p}_{ik} - \hat{p}_{T,ik}), \quad (16)$$

where $\hat{g}_{ik} = \hat{p}_{T,ik}(1 - \hat{p}_{T,ik})\hat{\sigma}_u^2(\hat{\sigma}_u^2\hat{p}_{T,ik}(1 - \hat{p}_{T,ik}) + \hat{\sigma}_{e,ik}^2)^{-1}$, and $\hat{\sigma}_{e,ik}^2$ is the LFS estimate of $\sigma_{e,ik}^2$. By construction, the univariate predictor in (16) is between the direct estimator and the corresponding SPREE estimator. However, the sum of the univariate predictors in a province may not equal 1, and the univariate predictors are not benchmarked to the direct estimators of the national three digit totals; $\sum_{k=1}^{K}\hat{p}_{pred,ik}M_{.k} \neq M_{i.}$. If $\hat{\sigma}_{e,ik}^2 = \hat{\sigma}_{a,k}^2\hat{p}_{T,ik}(1 - \hat{p}_{T,ik})$ for a constant $\hat{\sigma}_{a,k}^2$, then the estimated weight $\hat{g}_{ik}$ is the same for all three digit codes in a province; $\hat{g}_{ik} = \hat{g}_k = \hat{\sigma}_u^2(\hat{\sigma}_u^2 + \hat{\sigma}_{a,k}^2)^{-1}$, and the univariate predictors, defined in (16), in a single province sum to 1.

We use a final raking operation to benchmark the predictors. First, calculate an initial predictor of each total, $\dot{M}_{pred,ik} = \hat{p}_{pred,ik}M_{.k}$. Second, rake the initial predictors of the totals to the direct estimators of the marginal totals. Let $\tilde{M}_{ik}$ be the total for three digit code i and province k in the raked table. Finally, convert the raked predictors of totals to proportions. The benchmarked predictor of the proportion is $\tilde{p}_{ik} = \dot{M}_{.k}^{-1}\tilde{M}_{ik}$.

4.4 Estimation of the Mean Squared Error

We define estimators of the mean squared errors (MSE's) of the predictors in analogy with the Prasad Rao (1990) approach. The estimator of $\boldsymbol{\Sigma}_{ee,k}$ in the LFS is the product of a multinomial covariance matrix and the average design effect. For consistency with the application, we define an estimator of the MSE under an assumption that $\hat{\boldsymbol{\Sigma}}_{ee,k} = \hat{\sigma}_{a,k}^2\hat{\boldsymbol{\Gamma}}_{uu,k}$, where $\hat{\boldsymbol{\Gamma}}_{uu,k}$ is the estimator of $\boldsymbol{\Gamma}_{uu,k}$ obtained by replacing the unknown proportions in (5) by the SPREE estimators.

Our estimator of the MSE of the proportion associated with three digit code i and province k is a sum of three terms:

$$\widetilde{MSE}_{ik} = g_{1,ik} + g_{2,ik} + 2g_{3,ik}. \quad (17)$$

The first term is $g_{1,ik} = (\hat{\sigma}_u^2 + \hat{\sigma}_{a,k}^2)^{-1}[\hat{\sigma}_u^2\hat{p}_{T,ik}(1 - \hat{p}_{T,ik})\hat{\sigma}_{a,k}^2]$, which accounts for the error in the predictor (15) constructed with the unknown parameters. The second term, $g_{2,ik}$, accounts for uncertainty due to estimation of $p_{T,ik}$. Specifically, $g_{2,ik} = (1 - \hat{g}_k)^2\ddot{V}_{spree,ik,ik}$, where $\ddot{V}_{spree,ik,ik}$ is an estimator of the variance of $\hat{p}_{T,ik}$. Based on the linear approximation in (8), an approximation for the variance of the vector of SPREE estimators is $\boldsymbol{V}_{spree} = \boldsymbol{W}_T'(\boldsymbol{\Sigma}_{dd22} + \sigma_u^2\boldsymbol{B}_{uu22})\boldsymbol{W}_T$, where $\boldsymbol{\Sigma}_{dd22}$ and $\sigma_u^2\boldsymbol{B}_{uu22}$ are the variances of $\boldsymbol{f}$ and $\boldsymbol{v}$, respectively. Then,

$\ddot{V}_{spree,ik,ik}$ estimates the diagonal element of $\boldsymbol{V}_{spree}$ associated with three digit code i and province k. The third term is $g_{3,ik} = \hat{p}_{T,ik}(1-\hat{p}_{T,ik})\hat{\sigma}_{a,k}^4(\hat{\sigma}_u^2 + \hat{\sigma}_{a,k}^2)^{-3}\hat{V}(\hat{\sigma}_u^2)$, where

$$\hat{V}(\hat{\sigma}_u^2) = \frac{\hat{\sigma}_b^t[2(\hat{\Sigma}_a+\hat{\Sigma}_b\hat{\xi})^2]^{-1}2(\hat{\Sigma}_a+\hat{\Sigma}_b\hat{\sigma}_u^2)^2[2(\hat{\Sigma}_a+\hat{\Sigma}_b\hat{\xi})^2]^{-1}\hat{\sigma}_b}{(\hat{\sigma}_b^t[2(\hat{\Sigma}_a+\hat{\Sigma}_b\hat{\xi})^2]^{-1}\hat{\sigma}_b)^2} \quad (18)$$

is an estimator of the variance of $\hat{\sigma}_u^2$ derived under normality. The third term, $g_{3,ik}$, estimates the variance of $\hat{g}_k$ and a bias of $g_{1,ik}$ due to use of the estimator of σ_u^2 instead of the true parameter. The estimator of the MSE in (17) ignores variability in LFS estimator of the sampling variance as well as covariances that are zero under an assumption of normality but can be nonzero in general. If the estimator of σ_u^2 is an average of estimators from d uncorrelated time points, then we replace $\hat{V}(\hat{\sigma}_u^2)$ in $g_{3,ik}$ with $d^{-2}\sum_{t=1}^{d}\hat{V}_t(\hat{\sigma}_u^2)$, where $\hat{V}_t(\hat{\sigma}_u^2)$ is the estimator defined in (18) from time point t. We obtain analogous estimators of the MSE's of totals; we omit the details.

The estimator of the MSE ignores the effect of the final raking operation. It is tempting to use a linear approximation to estimate the contribution of raking to the MSE. One justification for the Taylor approximation for the raked predictors is that the small area effects converge to zero. Because we do not assume that the small area effects converge to zero, we do not use the linear approximation to the final raking operation in the estimator of the MSE.

5. SIMULATION

This section has three parts. First, we explain how we use the Dirichlet-multinomial distribution to generate variables to represent the direct estimators. Second, we specify estimators of the variances used in the simulation. Third, we describe the results of the Monte Carlo experiment.

5.1 Hierarchical Simulation Model

The data for the simulation study were generated to represent the two digit code A1 (specialist managers), cross-classified into 4 three digit codes by 10 provinces. The two-way table of totals from the 2006 Census of Canada and estimated province totals from a 2008 LFS were used to define the parameters. The expected values of the direct estimators of the national three digit totals were generated so that their sum is equal to the sum of the expected values of the province two digit totals and their order is concordant with the ordering in the 2006 Census. The expected values of the cell totals are the results of raking the Census totals to the fixed margins. The resulting expected values of the proportions are in Table 1, along with the fixed marginal totals.

Table 1: Expected Values of Proportions and Marginal Totals

Province	PE	NF	NS	NB	SK	MB	AB	BC	QC	ON	Three Digit Total, $T_{i.}$
$i=1$	0.401	0.297	0.284	0.317	0.300	0.298	0.280	0.270	0.363	0.277	99611
$i=2$	0.193	0.182	0.198	0.217	0.170	0.173	0.229	0.200	0.240	0.250	76368
$i=3$	0.243	0.259	0.274	0.264	0.265	0.298	0.278	0.319	0.270	0.301	96290
$i=4$	0.163	0.261	0.244	0.202	0.265	0.231	0.213	0.211	0.128	0.172	59766
Two Digit Total, $T_{.k}$	1398	5079	7473	9383	10160	12080	35851	43816	77505	129290	

To induce sampling variability in the direct estimators of the marginal totals, we generate $\tilde{n}_k$, the sample size for province k, from a binomial distribution with a sample size of m_k and a success probability equal to 0.01. The direct estimator of the province two digit total is $\hat{M}_{.k} = (m_k 0.01)^{-1}T_{.k}\tilde{n}_k$. We choose the expected sample size, $n_k = 0.01m_k$, to produce the coefficients of variation of the direct estimators of the province two digit totals in Table 2.

Table 2: Expected Province Sample Sizes and Coefficients of Variation of Direct Estimators of Province Margins

Province	PE	NF	NS	NB	SK	MB	AB	BC	QC	ON
Expected Sample Size, n_k	20.8	25.2	26	32	52	64.8	66	70	76	259
CV of $\hat{M}_{.k}$	0.209	0.190	0.186	0.168	0.131	0.118	0.117	0.113	0.109	0.059

Given the random province sample size, $\tilde{n}_k$, we generate the direct estimators of the proportions from Dirichlet-multinomial distributions. We first generate the vector of true proportions, $\boldsymbol{p}_k$, from a Dirichlet distribution with probability density function,

$$f(x_{1k}, \ldots, x_{Ik}) \propto \prod_{i=1}^{I} (x_{ik})^{\omega_{ik}-1}, \tag{19}$$

where $0 \leq x_{ik} \leq 1$, $\sum_{i=1}^{I} x_{ik} = 1$, and $\omega_{ik} = p_{ik}((\sigma_u^2)^{-1} - 1)$. We set $\sigma_u^2 = 0.01726$, the median of the inverses of the expected sample sizes. Given $\tilde{n}_k$ and $\boldsymbol{p}_k$, we generate $(\tilde{M}_{1k}^{(d)}, \ldots, \tilde{M}_{Ik}^{(d)})'$ from a multinomial distribution with probability mass function,

$$P(\tilde{M}_{1k}^{(d)} = m_1, \ldots, \tilde{M}_{Ik}^{(d)} = m_I) = \tilde{n}_k! (\prod_{i=1}^{I} m_i!)^{-1} \prod_{i=1}^{I} (p_{ik})^{m_i}, \tag{20}$$

where $m_1, \ldots, m_I$ are non-negative integers that sum to $\tilde{n}_k$. Finally, we compute the direct estimator of the proportion as $\hat{p}_{ik} = \tilde{n}_k^{-1} \tilde{M}_{ik}^{(d)}$ and the corresponding total as $\hat{M}_{ik} = \tilde{M}_{.k} \hat{p}_{ik}$.

To express the model for simulation in the form of (1), let $e_{ik} = \hat{p}_{ik} - p_{ik}$, and let $u_{ik} = p_{ik} - p_{T,ik}$. The moments of e_{ik} and u_{ik} depend on the moments of the multinomial and Dirichlet distributions. Under the simulation model, e_{ik} and u_{ik} have means of zero and are uncorrelated. The variance of $\boldsymbol{u}_k$ has the form in (5), and the variance of e_{ik} is

$$V\{e_{ik}\} = E[\tilde{n}_k^{-1} p_{ik}(1 - p_{ik})(1 - \sigma_u^2)]. \tag{21}$$

5.2 Variance Estimation in the Simulation

We define estimators of the sampling variances and the small area effects in the simulation to approximate the procedure used in the LFS. We use

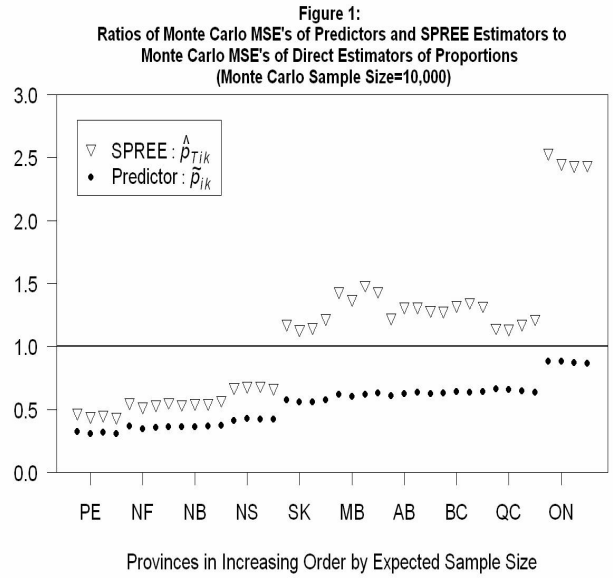
$$\hat{\boldsymbol{\Sigma}}_{ee,k} = \tilde{n}_k^{-1} \left[\mathrm{diag}(\hat{\boldsymbol{p}}_{T,k}) - \hat{\boldsymbol{p}}_{T,k} \hat{\boldsymbol{p}}_{T,k}' \right] \tag{22}$$

as the estimator of the variance of the vector of sampling errors in province k. The LFS estimator of σ_u^2 is smoothed across several time points, $t = 1, \ldots, d$. In the simulation, we use the procedure described in section 4.2 to combine estimators of σ_u^2 from four independently generated data sets.

5.3 Results

Ratios of the Monte Carlo MSE's of the predictors and the SPREE estimators to the Monte Carlo MSE's of the direct estimators of the proportions are plotted in Figure 1. The Monte Carlo sample size is 10,000. The provinces, on the horizontal axis, are listed in increasing order by the expected sample size, n_k. The expected sample sizes are in Table 2. In the three smallest provinces (PE, NF, and NB), the MSE's of the predictors are between 30.54% and 37.17% of the MSE's of the direct estimators (Monte Carlo standard errors of the ratios range from .57% to .65%), and the MSE's of the SPREE estimators are between 44.23% and 57.5% of the MSE's of the direct estimators (SE 0.93%-1.1%). In the province with the fourth smallest expected sample size (NS), the ratios of the MSE's of the predictors to the MSE's of the direct estimators are 41.09% to 42.61% (SE 0.69%-0.71%), while corresponding ratios for the SPREE estimators are between 67.0% and 69.0% (SE 1.4%). In the next five provinces (SK, MB, AB, BC, and QC), the MSE's of the predictors range from 55.57% to 65.40% of the MSE's of the direct estimators (SE 0.69% - 0.93%), and the MSE's of the SPREE estimators are 113.7% to 148.9% of the MSE's of the direct estimators (SE 2.2%-3.0%). In the province with the largest expected sample size (ON), the relative MSE's of the predictors are 86.52% to 88.33% (SE .83%-.87%), while the bias of the SPREE estimator leads to relative MSE's between 244.2% and 253.9% (SE 4.7%-4.9%).

Figure 2 shows relative MSE's of the predictors of totals and proportions. The relative MSE's of totals and proportions tend to increase as the expected sample sizes increase. However, the relative MSE's of totals vary more within each province than the relative MSE's of proportions. For example, in PE, the relative MSE's for totals range from 41.90% to 60.78% (SE 0.71%-0.85%), while the relative MSE's for proportions are between 30.54% and 32.05% (SE 0.57% - 0.59%). In each province, the reduction in the MSE is smallest for the cell with the largest expected value. For example, in PE, the largest of the expected values of the proportions is 0.401 (Table 1), and the relative MSE of the associated total is 60.78% (SE 0.85%). Also, Figure 2 shows that the reduction in the MSE due to the small area procedure is uniformly smaller for totals than for proportions. We conjecture that variability in the direct estimators of the province two digit totals accounts for the differences between the MSE's of totals and proportions. In particular, variability in the direct estimators of the province margins limits the possible reduction in the MSE's of totals.



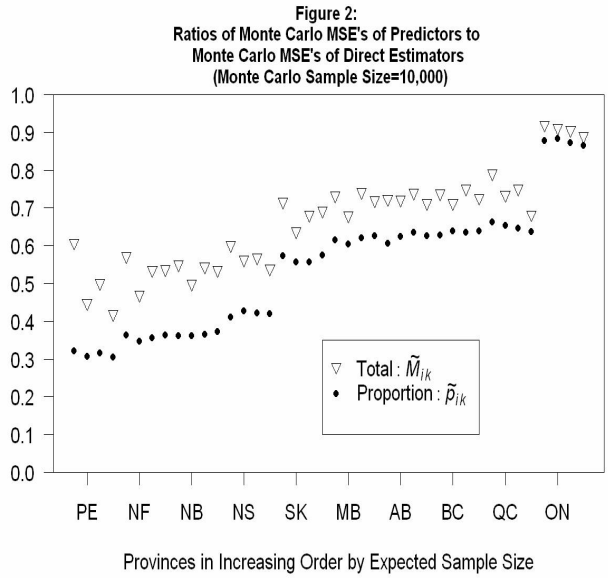
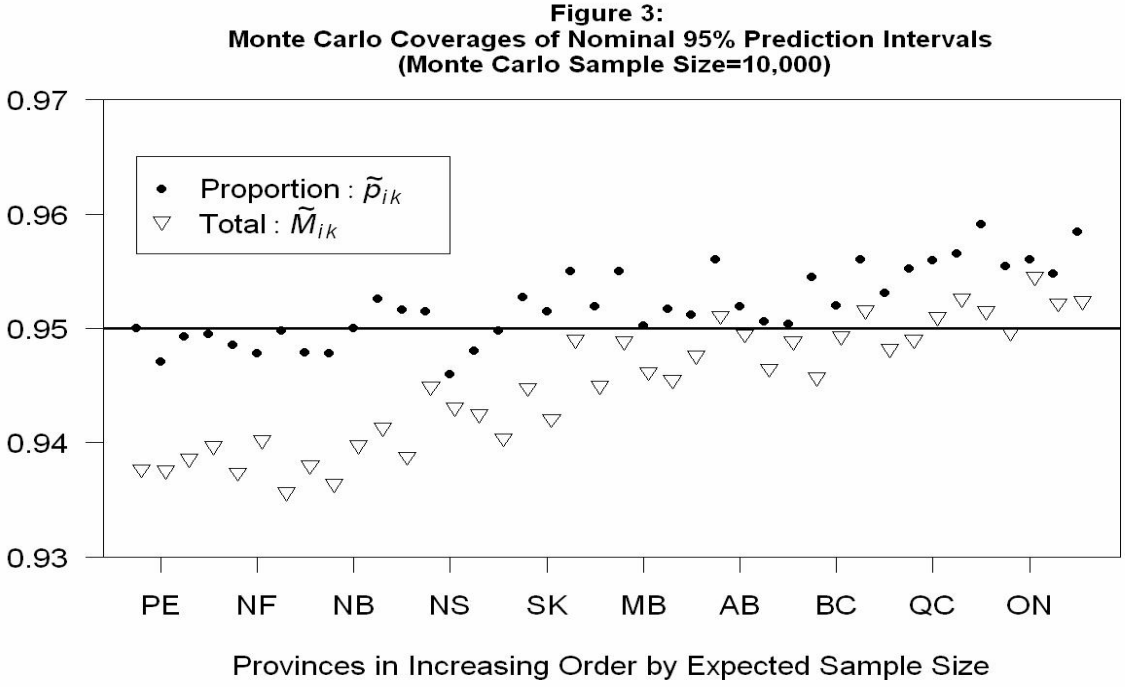


Figure 3 shows the empirical coverages of nominal 95% prediction intervals. The empirical coverages for proportions are close to 95% in the smaller provinces and are between 95% and 96% in the larger provinces. The empirical coverage for each cell total is smaller than the coverage for the corresponding proportion. For both totals and proportions, the empirical coverages tend to increase with the expected sample sizes. We do not have an explanation for this increasing trend.



6. SUMMARY

We develop a model based procedure that improves direct estimators of the cells of a two-way table when the margins of the table are precisely estimated. The procedure preserves the direct estimators of the marginal totals and incorporates information from a census. In a simulation, the MSE's of the predictors of totals and proportions are uniformly smaller than the MSE's of the corresponding direct estimators, and the relative MSE tends towards one as the expected sample size in a province increases. Due to variability in the direct estimators of the marginal totals, the reduction in the MSE due to the small area procedure is greater for proportions than for totals. Empirical coverages of nominal 95% prediction intervals are between 93% and 95% in the small provinces and are between 94% and 96% in the larger provinces.

REFERENCES

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.

Ghosh, M. and Lahiri, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, **82**, 1153-1162.

Griffiths, R. (1996). Current population survey small area estimation for congressional districts. *Proceedings of the American Statistical Association*, Section on Survey Research Methods, pp. 314-319.

Hidiroglou, M.A. and Patak Z. (2009). Small area estimation feasibility study for human resources and skills development Canada. Internal Statistical Research and Innovation Division working paper.

Isaki, C.T., Tsay, J.H., and Fuller, W.A. (2000). Estimation of Census adjustment factors. *Survey Methodology*, **26**, 31-42.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15, 1-96.

Liu, B., Lahiri, P., Kalton, G. (2007). Hierarchical Bayes modeling of survey-weighted small area proportions. *Proceedings of the American Statistical Association*, Section on Survey Research Methods, pp. 3181-3186.

Longford, N.T. (1999). Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society A*, **162**, 227-245.

Noble, A., Haslett, S., and Arnold, G. (2002). Small area estimation via generalized linear models. *Journal of Official Statistics*, **18**, 45-60.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163-171.

Purcell, N.J. and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, **48**, 3-18.

Raghunathan, T.E. (1993). A quasi-empirical Bayes method for small area estimation. *Journal of the American Statistical Association*, **88**, 1444-1448.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley.

Wang, J., Fuller, W.A. (2003). The mean squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, **98**, 716-723.

Zhang L. and Chambers R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society B*, **66**, 479-496.