A Small Area Procedure For Estimating Population Counts

Emily Berg[*] Wayne A. Fuller [†]

Abstract

When the cells of a contingency table are unplanned domains, small realized sample sizes can cause direct survey estimators to be unreliable. We develop a procedure that obtains more stable estimators of the cell totals and proportions in a two-way table under an assumption that the margins of the table are well estimated. The method preserves the direct estimators of the marginal totals and incorporates information from a previous census. The procedure was developed for the Canadian Labour Force Survey as a way to improve occupational detail at the province level.

In a simulation, the predictor achieves a smaller mean squared error than the direct estimator. Due to variability in the direct estimators of the marginal totals, the reduction in the MSE is greater for proportions than for totals. Empirical coverages of nominal 95% prediction intervals are betweeen 93% and 96%.

Key Words: Small Area Estimation, Benchmarking, Generalized Linear Structural Model

1. Introduction

Small area estimation is a term that describes estimation for domains in which realized sample sizes are too small to produce stable direct estimators. Small areas are often defined by a cross-classification of demographic and geographic variables. A widely adopted solution uses models to combine direct estimators with synthetic estimators.

Models for small area estimation typically have a hierarchical structure in which a "sampling model" describes the distributions of the direct estimators given the true values, and a "linking model" (Liu et al., 2007) relates 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 and non-normal error distributions. Rao (2003) and Jiang and Lahiri (2006) review methods for small area estimation.

The small area procedure that we suggest was developed for the Canadian Labour Force Survey (LFS). The objective is to obtain more reliable estimates of the cells in the two-way table defined by occupations and provinces. The direct estimators of the marginal totals are judged to have adequate design properties, and the two-way table from the previous Census provides auxiliary information.

Purcell and Kish (1980) propose Structure Preserving Estimation (SPREE) as a way to update the contingency table from a previous census using direct estimators of the marginal totals obtained from a current survey. SPREE preserves the interactions in the Census and the margins from the current survey. If the interactions in the census do not hold for the current time period, then the SPREE estimators are biased. Zhang and Chambers (2004) generalize the loglinear model underlying SPREE to a generalized linear model with interactions proportional to

[*] Center for Survey Statistics and Methodology, Department of Statistics, Iowa State University

[†] Center for Survey Statistics and Methodology, Department of Statistics, Iowa State University

the Census interactions. They then extend the generalized linear model to a generalized linear mixed model with normally distributed random effects. We suggest an alternative way to extend the generalized linear model to a nonlinear mixed model.

We organize the rest of this document as follows. We provide more background on the LFS application in section 2. We discuss the SPREE procedure and extensions of SPREE in more detail in section 3. We specify a model for the direct LFS estimators in section 4 and develop a model based predictor in section 5. We define an estimator of the MSE in section 5 and evaluate the procedure through simulation in section 6.

2. Canadian Labour Force Survey

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 three digit codes are nested in two digit codes. For example, the two digit code A1 denotes the category for specialist managers. The four three digit codes A11-A14 subdivide specialist managers into more specific occupations. (Hidiroglou and Patak, 2009)

The LFS estimation system provides weighted direct estimators of occupational totals through the three digit level of detail in each province. 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. Stable monthly estimators of three digit totals and proportions at the province level are desired. (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 the best available source of auxiliary information for estimating three digit occupational totals and proportions at the province level with the LFS data. (Hidiroglou and Patak, 2009)

Our objective is to obtain estimators of the cell totals in the two-way table defined by the cross-classification of three digit codes and provinces in a single two digit code with better precisions than the direct estimators. Because the direct LFS estimators of the margins of the two-way table have small coefficients of variation, we desire a predictor of the cell totals that preserves 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 employment totals in the current time point.

3. SPREE and Generalizations

Structure Preserving Estimation (SPREE) is a synthetic small area procedure that combines auxiliary information, often data from a previous census, with current survey data to improve the precision of estimators of the cell totals in a multi-way contingency table (Purcell and Kish, 1980). The idea underlying SPREE is that relationships inherent in the previous census serve as a good model for the current time period, while estimates of the marginal levels should come from the current survey because the census totals are out-dated. SPREE adjusts the interior of the table from the previous census in a way that preserves the interactions from the census and the margins from the current survey. Purcell and Kish (1980) implement SPREE by applying iterative proportional fitting to the two-way table with the census totals in the interior cells and the margins estimated from the current survey.

Noble et al. (2002) characterize the model underpinning SPREE as a special case of a generalized linear model. The estimators of the cell totals obtained from SPREE are the maximum likelihood estimators of the expected counts under a generalized linear model with a Poisson random component and a log link. Main effects for rows and columns are estimated with the direct estimators. Interactions are set equal to the interactions in a saturated loglinear model fit to the census two-way table. The representation of SPREE as maximum likelihood estimation suggests Newton-Raphson as an alternative to iterative proportional fitting as a way implement SPREE (Noble et al., 2002).

Noble et al. (2002) extend the SPREE model to the family of generalized linear models. In the general setting, the parameters of the linear predictor are partitioned into two sets: one set (eg., the main effects in the case of SPREE) is estimated from the direct estimators and the second set (eg., the interactions) from the auxiliary data. The representation of SPREE as a special case of a generalized linear model reveals that estimating a subset of the model parameters from an auxiliary source is not limited to categorical responses and explanatory variables. Noble et al. (2002) illustrate the generalization of SPREE through an application to estimation of unemployment rates from the Household Labour Force Survey conducted by Statistics New Zealand.

Griffiths (1996) considers two composite estimators of the cell totals in contingency tables defined by economic characteristics in congressional districts in Iowa. One of the composite estimators is a convex combination of the SPREE estimator and the corresponding direct estimator. The weights used to form the convex combination depend on the design MSE's of the SPREE estimators and the design MSE's of the direct estimators. The other composite estimator is the EBLUP based on a mixed linear model for the direct estimators of totals.

Zhang and Chambers (2004) develop two extensions of the loglinear model underlying SPREE. First, the generalized linear structural model (GLSM) permits the interactions from the current time point to be proportional to (rather than equal to, as in SPREE) the interactions from a census. A further extension of the GLSM to the generalized linear structural mixed model (GLSMM) incorporates random small area effects.

4. A Model for Three Digit Codes in Provinces

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 $i = 1, \ldots, m$, and $k = 1, \ldots, K$. We assume that the direct estimator of the proportion satisfies

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

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

$$p_{ik} = g_{ik}(\boldsymbol{N}, \boldsymbol{\lambda}_o) \tag{2}$$

is a function of the vector of Census totals $\boldsymbol{N}$ and a vector of parameters $\boldsymbol{\lambda}_o$. Assume u_{ik} and e_{jt} are uncorrelated for all i, k, j, t. Because we treat the Census totals as fixed, p_{ik} is a fixed parameter. The true cell proportion to be predicted is

$$p^*_{ik} = p_{ik} + u_{ik}. \tag{3}$$

The model (1) expresses the direct estimator as a sum of three parts: the marginal expected value, p_{ik}, the random small area effect, u_{ik}, and the random sampling error, e_{ik}. We discuss the specific assumptions about the form of the function (2) and the variances of the random components in the following subsections. We then discuss the implications of the model assumptions on the first and second moments of the direct estimators of the cell totals.

4.1 Fixed Expected Value

Let $(\alpha\beta)^{cen}_{ik}$ be the maximum likelihood estimate of the interaction in a saturated loglinear model that specifies the Census totals to have independent Poisson distributions with means $\{\mu^{cen}_{ik} : i = 1, \ldots, m, \text{ and } k = 1, \ldots, K\}$ satisfying

$$\log(\mu^{cen}_{ik}) = \alpha^{cen}_i + \beta^{cen}_k + (\alpha\beta)^{cen}_{ik}.$$

For estimability, set $\alpha^{cen}_1 = (\alpha\beta)^{cen}_{i1} = (\alpha\beta)^{cen}_{1k} = 0$ for $i = 1, \ldots, m$ and $k = 1, \ldots, K$. Define

$$T_{ik}(\boldsymbol{\lambda}) = \exp(\alpha_i + \beta_k + \theta(\alpha\beta)^{cen}_{ik}), \tag{4}$$

where $\alpha_1 = 0$, and $\boldsymbol{\lambda} = (\alpha_2, \ldots, \alpha_m, \beta_1, \ldots, \beta_K)'$. Then, assume the function in (2) is

$$p_{ik} = g_{ik}(\boldsymbol{N}, \boldsymbol{\lambda}_o) = T^{-1}_{.k} T_{ik},$$

where $\boldsymbol{\lambda}_o = (\alpha_{o,2}, \ldots, \alpha_{o,m}, \beta_{o,1}, \ldots, \beta_{o,K}, \theta_o)'$, $T_{ik} = T_{ik}(\boldsymbol{\lambda}_o)$, and $T_{.k} = \sum_{i=1}^m T_{ik}$. By construction, $\sum_{i=1}^m p_{ik} = 1$.

The loglinear model in (4) is the Generalized Linear Structural Model (GLSM) introduced by Zhang and Chambers (2004). The assumption that $\theta_o = 1$ produces the loglinear model underlying SPREE. The SPREE model specifies the odds ratios in the two-way table with T_{ik} in the cell for three digit code i and province k to equal the odds ratios in the two-way table of Census totals. Allowing θ_o to differ from 1 relaxes the assumption that the interactions in the Census persist unchanged through time.

In section 5, we estimate $\boldsymbol{\lambda}_o$ using the estimator that would be the maximum likelihood estimator under an assumption that the direct estimators of the totals are independent Poisson random variables. We call the resulting estimators of T_{ik} and p_{ik} the GLSM estimators. The GLSM estimators are equal to the SPREE estimators if the estimator of θ_o is constrained to equal 1. If the interactions in the table of the expected values of the direct estimators are proportional to the interactions in the Census, with a proportionality constant not equal to 1, then estimating θ_o updates the SPREE estimators so that the estimated interactions better represent the interactions in the current time point.

4.2 Small Area Effects

Let $\boldsymbol{u}_k = (u_{1k}, \ldots, u_{mk})'$ denote the vector of small area effects for a province. Assume that $E\{u_{ik}\} = 0$ and that the population covariance matrix for $\boldsymbol{u}_k$ is

$$\boldsymbol{\Sigma}_{uu,k} = \psi[\text{diag}(\boldsymbol{p}_k) - \boldsymbol{p}_k\boldsymbol{p}_k'] := \psi\boldsymbol{\Gamma}_{uu,k}, \tag{5}$$

where $\boldsymbol{p}_k = (p_{1k}, \ldots, p_{mk})'$, and ψ is a constant. The variance of $\boldsymbol{u}_k$ in (5) is proportional to the multinomial covariance matrix. Assuming the covariance between $\boldsymbol{u}_k$ and $\boldsymbol{u}_t$ is zero for $t \neq k$, the covariance matrix of the vector of small area effects, $\boldsymbol{u} = (\boldsymbol{u}_1', \ldots, \boldsymbol{u}_k')'$, is block-diagonal with $\psi\boldsymbol{\Gamma}_{uu,k}$ as the k^{th} block. Because the columns of the covariance matrix in (5) sum to zero, $\sum_{i=1}^m u_{ik} = 0$ for all k.

4.3 Sampling Errors

Assume that $E[e_{ik} \mid u_{ik}] = 0$ so that the direct estimators are conditionally unbiased for the true proportions defined in (3), given u_{ik}. Let $\boldsymbol{\Sigma}_{ee,k}$ denote the variance of $\boldsymbol{e}_k = (e_{1k}, \ldots, e_{mk})'$. In the LFS, sampling is independent across provinces, so $\boldsymbol{\Sigma}_{ee}$, the covariance matrix of the vector of design errors, $\boldsymbol{e} = (\boldsymbol{e}_1', \ldots, \boldsymbol{e}_K')'$, is block-diagonal with $\boldsymbol{\Sigma}_{ee,k}$ as the k^{th} block. Because the LFS estimators of the proportions in a single province sum to 1, the covariance matrix for $\boldsymbol{e}_k$ is such that $\sum_{i=1}^m e_{ik} = 0$ for all k.

4.4 True Totals and Corresponding Direct Estimators

The true total in three digit code i and province k is

$$M_{ik}^* = (p_{ik} + u_{ik})T_{.k}. \tag{6}$$

Because $p_{ik}T_{.k} = T_{ik}$, T_{ik} is the expected value of M_{ik}^*. Because $\sum_{i=1}^m u_{ik} = 0$ and $\sum_{i=1}^m p_{ik} = 1$, the true province two digit total is equal to its expected value;

$$\sum_{i=1}^m M_{ik}^* := M_{.k} = T_{.k}.$$

In contrast, the expected value of the true national three digit total is a function of the u_{ik}. The true national three digit total is

$$M_{i.} = \sum_{k=1}^K M_{ik}^* = \sum_{k=1}^K (p_{ik} + u_{ik})T_{.k}.$$

The expected value of $M_{i.}$ is

$$E[M_{i.}] = \sum_{k=1}^K p_{ik}T_{.k} := T_{i.}.$$

Because

$$\sum_{k=1}^K T_{.k}u_{ik} \neq 0, \tag{7}$$

$M_{i.}$ does not equal $T_{i.}$. However, under the assumption that $0 < c_1 < T_{.k} < c_2 < \infty$ and that $\boldsymbol{u}$ has sufficient moments, $K^{-1}M_{i.} - K^{-1}T_{i.}$ converges to zero in probability as K increases.

The model for the proportions has implications for the first and second moments of the direct estimators of the cell totals. Let $\hat{M}_{.k}$ be the direct estimator of the two digit total in province k. Assume $\hat{M}_{.k}$ is unbiased for $T_{.k}$ and is independent of the vector of small area effects $\boldsymbol{u}$. Because the direct estimator of the total and the proportion are related through the identity, $\hat{M}_{ik} = \hat{M}_{.k}\hat{p}_{ik}$, it follows that

$$\hat{M}_{ik} \quad = \quad T_{ik} + u_{ik}\hat{M}_{.k} + a_{ik}, \tag{8}$$

where $a_{ik} = e_{ik}\hat{M}_{.k} + p_{ik}(\hat{M}_{.k} - T_{.k})$, and a_{ik} is the sampling error in the scale of totals. By (8) and the assumption that u_{ik} and $\hat{M}_{.k}$ are uncorrelated, $E\{\hat{M}_{ik}\} = T_{ik}$.

Let $(\boldsymbol{u}', \boldsymbol{u}'_{\hat{M}})'$ be the vector of small area effects, where $\boldsymbol{u}_{\hat{M}} = (\boldsymbol{u}'_{\hat{M},1}, \ldots, \boldsymbol{u}'_{\hat{M},K})'$, and $\boldsymbol{u}_{\hat{M},k} = \boldsymbol{u}_k\hat{M}_{.k}$. By the assumption that $\hat{M}_{.k}$ is independent of u_{ik} (for all i, k), the variance of $(\boldsymbol{u}', \boldsymbol{u}'_{\hat{M}})'$ has the form,

$$V\{(\boldsymbol{u}', \boldsymbol{u}'_{\hat{M}})'\} = \psi \boldsymbol{B}_{uu},$$

where $\boldsymbol{B}_{uu}$ is a function of $\boldsymbol{\Gamma}_{uu}$ and of the expected values of the province two digit totals. Let $\boldsymbol{\Sigma}_{dd}$ denote the large sample variance of the vector of sampling errors $(\boldsymbol{e}', \boldsymbol{a}')'$, where $\boldsymbol{a} = (\boldsymbol{a}'_1, \ldots, \boldsymbol{a}'_K)'$, and $\boldsymbol{a}_k = (a_{1k}, \ldots, a_{mk})'$. A Taylor expansion can be used to express $\boldsymbol{\Sigma}_{dd}$ as a function of the expected values of the direct estimators and the variance of $\boldsymbol{e}$. The specific form of $\boldsymbol{B}_{uu}$ and a description of the Taylor expansion used to derive $\boldsymbol{\Sigma}_{dd}$ are omitted.

To summarize, the quantities of interest are the true proportions, defined in (3), and the true totals, defined in (6). The data available for prediction include the direct estimators of the proportions, defined in (1), and the corresponding estimators of the cell totals, defined in (8). The Census table provides auxiliary information.

5. Procedure

The suggested procedure for predicting the true totals, defined in (6), and the proportions, defined in (3), is composed of several steps. First, we estimate the fixed expected value, p_{ik}, using the Poisson score function as a set of estimating equations. Second, we obtain an estimator of ψ, defined in (5), using approximations for expected mean squares. Third, we calculate an initial predictor as a weighted combination of the direct estimator and of the estimator of p_{ik} from the first step, with weights determined by estimators of the variances. Because the initial predictors are not calibrated to the direct estimators of the marginal totals, we use a final raking operation to benchmark the predictors. Finally, we use a linear approximation to define an estimator of the MSE.

5.1 Estimator of Fixed Expected Value

Following Noble et al. (2002) and Zhang and Chambers (2004), we use the Poisson score function to estimate the parameter vector $\boldsymbol{\lambda}_o$, defined following (4). The resulting estimators of the expected values of the totals satisfy the restrictions that the sum across provinces for a fixed three digit code equals the direct estimator of the national three digit total, and the sum across the three digit codes in a province equals the direct estimator of the province two digit total. The score function under a model that specifies the direct estimators of the totals to be independent Poisson random variables with means $\{T_{ik} : i = 1, \ldots, m, \text{ and } k = 1, \ldots, K\}$, defined following (4), is

$$\boldsymbol{s}(\boldsymbol{\lambda}) = \boldsymbol{X}'(\hat{\boldsymbol{M}} - \boldsymbol{T}(\boldsymbol{\lambda})), \tag{9}$$

where $\boldsymbol{X}$ is the $mK \times (m+K)$ matrix with $\boldsymbol{x}'_{ik} = (I[i=2], \ldots, I[i=m], I[k=1], \ldots, I[k=K], (\alpha\beta)^{cen}_{ik})$ in row $m(k-1)+i$ (the row corresponding to three digit code i and province k), $I[\cdot]$ is the indicator function, $\boldsymbol{T}(\boldsymbol{\lambda}) = (\boldsymbol{T}_1(\boldsymbol{\lambda})', \ldots, \boldsymbol{T}_K(\boldsymbol{\lambda})')'$, $\boldsymbol{T}_k(\boldsymbol{\lambda}) = (T_{1k}(\boldsymbol{\lambda}), \ldots, T_{mk}(\boldsymbol{\lambda}))'$, and $\hat{\boldsymbol{M}}$ is the vector of direct estimators of cell totals listed in the order corresponding to the order of the elements of $\boldsymbol{T}(\boldsymbol{\lambda})$. The estimator $\hat{\boldsymbol{\lambda}} = (\hat{\alpha}_2, \ldots, \hat{\alpha}_m, \hat{\beta}_1, \ldots, \hat{\beta}_K, \hat{\theta})'$ of $\boldsymbol{\lambda}_o$ satisfies $\boldsymbol{s}(\hat{\boldsymbol{\lambda}}) = \boldsymbol{0}$. The estimator of the expected value of the proportion is

$$\hat{p}_{T,ik} = \hat{M}^{-1}_{.k}\hat{T}_{ik}, \tag{10}$$

where $\hat{T}_{ik} = T_{ik}(\hat{\boldsymbol{\lambda}}) = \exp(\hat{\alpha}_i + \hat{\beta}_k + \hat{\theta}(\alpha\beta)^{cen}_{ik})$, and $\hat{\alpha}_1 = 0$. As discussed in section 4, we refer to $\hat{T}_{ik}$ and $\hat{p}_{T,ik}$ as the GLSM estimators. The GLSM estimators satisfy the marginal restrictions; $\sum_{i=1}^{m} \hat{T}_{ik} = \hat{M}_{.k}$, and $\sum_{k=1}^{K} \hat{T}_{ik} = \hat{M}_{i.}$, where $\hat{M}_{i.} = \sum_{k=1}^{K} \hat{M}_{ik}$.

The estimator $\hat{p}_{T,ik}$ is approximately linear in the direct estimators of the cell totals. To justify the linear approximation, we assume that $\hat{\boldsymbol{\lambda}}$ converges in probability to $\boldsymbol{\lambda}_o$ as the dimensions of the two-way table and the province sample sizes increase. By a Taylor expansion of the score function in (9) around the true parameter $\boldsymbol{\lambda}_o$,

$$\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_o \;\approx\; (\boldsymbol{X}'\mathrm{diag}(\boldsymbol{T})\boldsymbol{X})^{-1}\boldsymbol{X}'(\hat{\boldsymbol{M}} - \boldsymbol{T}), \tag{11}$$

where $\boldsymbol{T} = (\boldsymbol{T}'_1, \ldots, \boldsymbol{T}'_K)'$, $\boldsymbol{T}_k = (T_{1k}, \ldots, T_{mk})'$, and $\mathrm{diag}(\boldsymbol{T})$ is the diagonal matrix with the vector $\boldsymbol{T}$ on the diagonal. To complete the linear approximation, let $\hat{\boldsymbol{p}}_T = (\hat{\boldsymbol{p}}'_{T,1}, \ldots, \hat{\boldsymbol{p}}'_{T,K})'$, where $\hat{\boldsymbol{p}}_{T,k} = (\hat{p}_{T,1k}, \ldots, \hat{p}_{T,mk})'$, and let $\boldsymbol{D}_p$ denote the matrix of derivatives of $\hat{\boldsymbol{p}}_T$ with respect to $\hat{\boldsymbol{\lambda}}$, evaluated at $\boldsymbol{\lambda}_o$. By (11),

$$\hat{\boldsymbol{p}}_T - \boldsymbol{p} \;\approx\; \boldsymbol{W}'_T(\boldsymbol{u}_{\hat{M}} + \boldsymbol{a}), \tag{12}$$

where $\boldsymbol{W}'_T = \boldsymbol{D}_p(\boldsymbol{X}'\mathrm{diag}(\boldsymbol{T})\boldsymbol{X})^{-1}\boldsymbol{X}'$, $\boldsymbol{p} = (\boldsymbol{p}'_1, \ldots, \boldsymbol{p}'_K)'$, and $\boldsymbol{u}_{\hat{M}} + \boldsymbol{a} = \hat{\boldsymbol{M}} - \boldsymbol{T}$ by (8).

5.2 Estimator of Model Variance

An estimated generalized least squares (EGLS) estimator of the variance of the small area random effects is constructed under an approximation for $E[(\hat{p}_{ik} - \hat{p}_{T,ik})^2]$. Wang and Fuller (2003) discuss the procedure that we use.

The EGLS estimator is based on a linear approximation for the difference, $\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T$, where $\hat{\boldsymbol{p}} = (\hat{\boldsymbol{p}}'_1, \ldots, \hat{\boldsymbol{p}}'_K)'$, and $\hat{\boldsymbol{p}}_k = (\hat{p}_{1k}, \ldots, \hat{p}_{mk})'$. Under our model (1) and by the Taylor expansion in (12),

$$\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T \approx \boldsymbol{W}'(\boldsymbol{e}', \boldsymbol{a}')' + \boldsymbol{W}'(\boldsymbol{u}', \boldsymbol{u}'_{\hat{M}})', \tag{13}$$

where $\boldsymbol{W}' = (\boldsymbol{I}_{mK}, \quad -\boldsymbol{W}'_T)$, and $\boldsymbol{I}_{mK}$ denotes the $mK \times mK$ identity matrix. Let $(\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T)^2$ be the $mK \times 1$ vector with elements that are the squares of the elements of $\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T$. Let $\boldsymbol{\Sigma}_a = \boldsymbol{W}'\boldsymbol{\Sigma}_{dd}\boldsymbol{W}$ and $\boldsymbol{\Sigma}_b = \boldsymbol{W}'\boldsymbol{B}_{uu}\boldsymbol{W}$, and let $\boldsymbol{\sigma}_a$ and $\boldsymbol{\sigma}_b$ be the vectors containing the diagonal elements of $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_b$, respectively. The matrices $\boldsymbol{W}$, $\boldsymbol{\Sigma}_{dd}$, and $\boldsymbol{B}_{uu}$ are functions of the expected values of the direct estimators and the sampling variances. We estimate $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_b$ by replacing the expected values with the GLSM estimators and the sampling variances with the LFS estimators of the design variances. Let $\hat{\boldsymbol{\Sigma}}_a$ and $\hat{\boldsymbol{\Sigma}}_b$ be the resulting estimates of $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_b$, and let $\hat{\boldsymbol{\sigma}}_a$ and $\hat{\boldsymbol{\sigma}}_b$ be the associated vectors of diagonal elements. By (13),

$$E[(\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T)^2] \approx \boldsymbol{\sigma}_a + \boldsymbol{\sigma}_b\psi, \tag{14}$$

and under an assumption of normality,

$$V\{(\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T)^2\} \approx 2(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b\psi)^2, \tag{15}$$

where the elements of the matrix $(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b\psi)^2$ are the squares of the elements of the matrix $(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b\psi)$.

An EGLS estimator of ψ under the linear model defined by (14) and (15) requires an initial estimate of ψ. We obtain an initial estimate of ψ from an estimator of the lower bound of the variance of an EGLS estimator. If $\psi = 0$, then, under an assumption of normally distributed errors, an estimator of the variance of an EGLS estimator is

$$\hat{V}_2(\tilde{\psi}_0) = [\hat{\boldsymbol{\sigma}}_b'(2\hat{\boldsymbol{\Sigma}}_a^2)^{-1}\hat{\boldsymbol{\sigma}}_b]^{-1}, \tag{16}$$

where $\hat{\boldsymbol{\Sigma}}_a^2$ contains 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 $\xi = 0.5[\hat{V}_2(\tilde{\psi}_0)]^{.5}$ as the initial estimate of ψ.

Setting $\psi = \xi$ in the variance expression (15), an EGLS estimator of ψ is

$$\tilde{\psi} = \frac{\hat{\boldsymbol{\sigma}}_b'[2(\hat{\boldsymbol{\Sigma}}_a + \hat{\boldsymbol{\Sigma}}_b\xi)^2]^{-1}((\hat{\boldsymbol{p}} - \hat{\boldsymbol{p}}_T)^2 - \hat{\boldsymbol{\sigma}}_a)}{\hat{\boldsymbol{\sigma}}_b'[2(\hat{\boldsymbol{\Sigma}}_a + \hat{\boldsymbol{\Sigma}}_b\xi)^2]^{-1}\hat{\boldsymbol{\sigma}}_b}. \tag{17}$$

The EGLS estimator (17) ignores the effect of replacing $\boldsymbol{\sigma}_a$ with an estimator. We use

$$\hat{\psi} = \max(\xi, \tilde{\psi}) \tag{18}$$

as the estimator of ψ. The initial value, ξ, imposes a strictly positive lower bound on the final estimator, $\hat{\psi}$.

If additional sources of information about ψ are available, one can improve the estimator of ψ. For example, the LFS survey supplies estimators of ψ from several time points, $t = 1, \ldots, d$. Let $\hat{V}_2^{(t)}$ denote the estimator (16) from the t^{th} time point, and let $\tilde{\psi}^{(t)}$ denote the corresponding EGLS estimator (17), which may be negative. Then, a smoothed estimator of ψ is

$$\hat{\psi}_d = \max(\bar{V}_2, \bar{\psi}), \tag{19}$$

where $\bar{V}_2 = (2d)^{-1}\left[\sum_{t=1}^{d}\hat{V}_2^{(t)}\right]^{.5}$, and $\bar{\psi} = d^{-1}\left[\sum_{t=1}^{d}\tilde{\psi}^{(t)}\right]$. Incorporating multiple time points has the potential to reduce the variance of the estimator of ψ and subsequently improve the precisions of the predictors.

5.3 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 MSE convex combination of $\hat{p}_{ik}$ and p_{ik}. We benchmark the initial predictors to the direct estimators of the marginal totals.

If we restrict to predictors that are linear in the direct estimators of the proportions, then a vector of minimum mean squared error predictors of the proportions in province k is $\boldsymbol{p}_{blp,k} = \boldsymbol{p}_k + \psi\boldsymbol{\Gamma}_{uu,k}(\boldsymbol{\Sigma}_{ee,k} + \psi\boldsymbol{\Gamma}_{uu,k})^{-}(\hat{\boldsymbol{p}}_k - \boldsymbol{p}_k)$, where "$-$" denotes a generalized inverse. Because the elements of $\boldsymbol{p}_{blp,k}$ may be larger than 1 or smaller than 0, we consider a univariate predictor of the form $\gamma_{ik}\hat{p}_{ik} + (1 - \gamma_{ik})p_{ik}$. The value of γ_{ik} that minimizes the mean squared error, $E[(\gamma_{ik}\hat{p}_{ik} + (1-\gamma_{ik})p_{ik} - p^*_{ik})^2]$, is

$$\gamma_{ik} = \frac{\psi p_{ik}(1 - p_{ik})}{\psi p_{ik}(1 - p_{ik}) + \sigma^2_{e,ik}}, \tag{20}$$

where $\sigma^2_{e,ik}$ is the i^{th} diagonal element of $\mathbf{\Sigma}_{ee,k}$. The associated predictor is

$$\tilde{p}_{ik}(\boldsymbol{p}, \boldsymbol{\gamma}) = p_{ik} + \gamma_{ik}(\hat{p}_{ik} - p_{ik}). \tag{21}$$

Isaki and Fuller (2000) compare univariate predictors calculated with the diagonal elements of an estimated covariance matrix with the empirical BLUP. In their simulation study, the univariate predictors have smaller MSE's than the predictors based on the full estimated covariance matrix.

If

$$\mathbf{\Sigma}_{ee,k} = \sigma^2_{a,k}\mathbf{\Gamma}_{uu,k} \tag{22}$$

for a constant $\sigma^2_{a,k}$, then the element of $\boldsymbol{p}_{blp,k}$ corresponding to three digit code i and province k is equal to $\tilde{p}_{ik}(\boldsymbol{p}, \boldsymbol{\gamma})$. Also, when (22) holds, the weight γ_{ik} is constant for all three digit codes in province k; $\gamma_{ik} = \gamma_k = \psi(\psi + \sigma^2_{a,k})^{-1}$.

The predictor (21) depends on unknown parameters. To calculate the predictor, we use the EGLS estimator of ψ and the GLSM estimator of p_{ik}. We compute an initial predictor for cell (i, k) as

$$\hat{p}_{pred,ik} = \hat{p}_{T,ik} + \hat{\gamma}_{ik}(\hat{p}_{ik} - \hat{p}_{T,ik}), \tag{23}$$

where

$$\hat{\gamma}_{ik} = \frac{\hat{\psi}\hat{p}_{T,ik}(1 - \hat{p}_{T,ik})}{\hat{\psi}\hat{p}_{T,ik}(1 - \hat{p}_{T,ik}) + \hat{\sigma}^2_{e,ik}}, \tag{24}$$

and $\hat{\sigma}^2_{e,ik}$ is the i^{th} diagonal element of an estimator of $\mathbf{\Sigma}_{ee,k}$. By construction, the univariate predictor in (23) is between the direct estimator and the GLSM estimator. An undesirable property of $\hat{p}_{pred,ik}$ is that the sum across the three digit codes in a single province may not equal 1. Also, the univariate predictors are not benchmarked to the direct estimators of the national three digit totals;

$$\sum_{k=1}^{K} \hat{p}_{pred,ik}\hat{M}_{.k} \neq \hat{M}_{i.}.$$

If the estimate of $\mathbf{\Sigma}_{ee,k}$ is proportional to the estimate of $\mathbf{\Gamma}_{uu,k}$, then the estimated coefficient (24) is the same for all three digit codes in a province;

$$\hat{\gamma}_{ik} = \hat{\gamma}_k = \frac{\hat{\psi}}{\hat{\psi} + \hat{\sigma}^2_{a,k}}, \tag{25}$$

where the estimate of $\mathbf{\Sigma}_{ee,k}$ is

$$\hat{\mathbf{\Sigma}}_{ee,k} = \hat{\sigma}^2_{a,k}\hat{\mathbf{\Gamma}}_{uu,k} = \hat{\sigma}^2_{a,k}[\text{diag}(\hat{\boldsymbol{p}}_{T,k}) - \hat{\boldsymbol{p}}_{T,k}\hat{\boldsymbol{p}}'_{T,k}]. \tag{26}$$

When (26) holds, the total of the predicted proportions in a province is 1 because $\sum_{i=1}^{m}(\hat{p}_{ik} - \hat{p}_{T,ik}) = 0$.

An initial estimator of the total employed in three digit code i and province k is

$$\hat{M}^*_{ik} = \hat{p}_{pred,ik}\hat{M}_{.k}. \tag{27}$$

Because the table with $\hat{M}^*_{ik}$ in the entry for three digit code i and province k does not necessarily satisfy the desired benchmarking property, we suggest using a final raking operation to benchmark the predictors. Let $\tilde{M}_{ik}$ denote the predicted total in three digit code i and province k in the raked table. Define the proportions arising from the benchmarked predictors of totals by

$$\tilde{p}_{ik} = \frac{\tilde{M}_{ik}}{\hat{M}_{.k}} \tag{28}$$

for $i = 1, \ldots, m$ and $k = 1, \ldots, K$. Unlike the initial univariate predictor $\hat{p}_{pred,ik}$, the benchmarked proportion in (28) is not restricted to fall between the direct estimator and the GLSM estimator. However, the benchmarked proportions satisfy the marginal restrictions; $\sum_{i=1}^{m} \tilde{p}_{ik}\hat{M}_{.k} = \hat{M}_{.k}$, and $\sum_{k=1}^{K} \tilde{p}_{ik}\hat{M}_{.k} = \hat{M}_{i.}$.

5.4 Estimator of Mean Squared Error

We use Taylor series to approximate the MSE of the predictor. Define

$$\tilde{p}_{ik}(\boldsymbol{\gamma}) = \hat{p}_{T,ik} + \gamma_{ik}(\hat{p}_{ik} - \hat{p}_{T,ik}), \tag{29}$$

which is the univariate predictor calculated with the unknown γ_{ik} defined in (20). Let $\tilde{\boldsymbol{p}}(\boldsymbol{\gamma})$ be the vector with element $\tilde{p}_{ik}(\boldsymbol{\gamma})$, let $\boldsymbol{D}_\gamma$ be the diagonal matrix with diagonal element γ_{ik}, and let $\boldsymbol{p}^*$ be the vector of true proportions defined in (3). The elements of $\tilde{\boldsymbol{p}}(\boldsymbol{\gamma})$, $\boldsymbol{D}_\gamma$, and $\boldsymbol{p}^*$ are listed in the order with provinces grouped together. By the linear approximations in the previous section,

$$\begin{aligned} \tilde{\boldsymbol{p}}(\boldsymbol{\gamma}) - \boldsymbol{p}^* &\approx \boldsymbol{D}_\gamma(\boldsymbol{u}+\boldsymbol{e}) - \boldsymbol{u} + (\boldsymbol{I}_{mK} - \boldsymbol{D}_\gamma)\boldsymbol{W}_T'(\boldsymbol{u}_{\hat{M}} + \boldsymbol{a}) \\ &= \boldsymbol{D}_1(\boldsymbol{e}', \boldsymbol{a}')' + \boldsymbol{D}_2(\boldsymbol{u}', \boldsymbol{u}_{\hat{M}}')', \end{aligned} \tag{30}$$

where $\boldsymbol{D}_1 = \begin{pmatrix} \boldsymbol{D}_\gamma, & (\boldsymbol{I}_{mK} - \boldsymbol{D}_\gamma)\boldsymbol{W}_T' \end{pmatrix}$ and $\boldsymbol{D}_2 = \begin{pmatrix} (\boldsymbol{D}_\gamma - \boldsymbol{I}_{mK}), & (\boldsymbol{I} - \boldsymbol{D}_\gamma)\boldsymbol{W}_T' \end{pmatrix}$. An approximation for the MSE based on (30) is

$$MSE_1 = \boldsymbol{D}_1\boldsymbol{\Sigma}_{dd}\boldsymbol{D}_1' + \boldsymbol{D}_2\psi\boldsymbol{B}_{uu}\boldsymbol{D}_2'. \tag{31}$$

If $\boldsymbol{\Sigma}_{ee,k} = \sigma^2_{a,k}\boldsymbol{\Gamma}_{uu,k}$, and if

$$E[\boldsymbol{D}_\gamma\boldsymbol{e}((\boldsymbol{u}_{\hat{M}} + \boldsymbol{a})'\boldsymbol{W}_T - \boldsymbol{u})(\boldsymbol{I}_{mK} - \boldsymbol{D}_\gamma)'] = \boldsymbol{0}, \tag{32}$$

then MSE_1 simplifies to

$$MSE_1 = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uu}(\boldsymbol{\Sigma}_{uu} + \boldsymbol{\Sigma}_{ee})^{-}\boldsymbol{\Sigma}_{uu} + (\boldsymbol{I}_{mK} - \boldsymbol{D}_\gamma)\boldsymbol{V}_{glsm}(\boldsymbol{I}_{mK} - \boldsymbol{D}_\gamma)', \tag{33}$$

where $\boldsymbol{\Sigma}_{uu}$ and $\boldsymbol{\Sigma}_{ee}$ are the block-diagonal covariance matrices of $\boldsymbol{u}$ and $\boldsymbol{e}$, respectively, and $\boldsymbol{V}_{glsm}$ is the approximate covariance matrix of $\hat{\boldsymbol{p}}_T$. Based on the linear approximation in (12), $\boldsymbol{V}_{glsm} = \boldsymbol{W}_T'(\psi\boldsymbol{B}_{uu22} + \boldsymbol{\Sigma}_{dd22})\boldsymbol{W}_T$, where $\boldsymbol{\Sigma}_{dd22}$ and $\boldsymbol{B}_{uu22}$ are the $mK \times mK$ sub-matrices of $\boldsymbol{\Sigma}_{dd}$ and $\boldsymbol{B}_{uu}$ giving the variances of $\boldsymbol{a}$ and $\boldsymbol{u}_{\hat{M}}$, respectively. An estimator of the diagonal element of the matrix in (33) corresponding to three digit code i and province k is

$$M\hat{S}E_{1,ik} \;=\; \frac{\hat{\psi}\hat{p}_{T,ik}(1-\hat{p}_{T,ik})\hat{\sigma}^2_{a,k}}{\hat{\psi} + \hat{\sigma}^2_{a,k}} + (1-\hat{\gamma}_k)^2\hat{V}_{glsm,ik,ik}, \tag{34}$$

where $\hat{V}_{glsm,ik,ik}$ denotes the estimator of the diagonal element of $\boldsymbol{V}_{glsm}$ corresponding to three digit code i and province k. The estimator (34) has the form

$$M\hat{S}E_{1,ik} = g_{1,ik}(\hat{\psi}) + g_{2,ik}(\hat{\psi}).$$

The first term in the sum is $g_{1,ik}(\hat{\psi}) = [\hat{\psi} + \hat{\sigma}^2_{a,k}]^{-1}[\hat{\psi}\hat{p}_{T,ik}(1-\hat{p}_{T,ik})\hat{\sigma}^2_{a,k}]$, which accounts for the prediction uncertainty due to the small area effects (the difference between $\tilde{p}_{ik}(\boldsymbol{p},\boldsymbol{\gamma})$ and p^*_{ik}). The second term is $g_{2,ik}(\hat{\psi}) = (1-\hat{\gamma}_k)^2\hat{V}_{glsm,ik,ik}$, which accounts for uncertainty due to estimation of p_{ik} (the difference between $\tilde{p}_{ik}(\boldsymbol{\gamma})$ and $\tilde{p}_{ik}(\boldsymbol{p},\boldsymbol{\gamma})$). The estimator (34) ignores variability in the predictor due to estimation of ψ.

We define an estimator that accounts for uncertainty due to estimation of ψ in analogy with the Prasad-Rao (1990) approach. Under normality, an estimator of the variance of $\hat{\psi}$ is

$$\hat{V}(\hat{\psi}) = \frac{\hat{\boldsymbol{\sigma}}_b'\boldsymbol{V}_\xi^{-1}\boldsymbol{V}_{\hat{\psi}}\boldsymbol{V}_\xi^{-1}\hat{\boldsymbol{\sigma}}_b}{(\hat{\boldsymbol{\sigma}}_b'\boldsymbol{V}_\xi^{-1}\hat{\boldsymbol{\sigma}}_b)^2}, \tag{35}$$

where $\boldsymbol{V}_\xi = 2(\hat{\boldsymbol{\Sigma}}_a + \hat{\boldsymbol{\Sigma}}_b\xi)^2$, and $\boldsymbol{V}_{\hat{\psi}} = 2(\hat{\boldsymbol{\Sigma}}_a + \hat{\boldsymbol{\Sigma}}_b\hat{\psi})^2$. If the estimator of ψ is an average of d estimators from d uncorrelated time points, as in (19), then an estimator of the variance of the pooled estimator of ψ is

$$\hat{V}_d = d^{-2}\sum_{t=1}^{d}\hat{V}(\hat{\psi}^{(t)}), \tag{36}$$

where $\hat{V}(\hat{\psi}^{(t)})$ is the estimator (35) from the t^{th} time point. Letting $\gamma_k'(\hat{\psi})$ denote the partial derivative of $\hat{\gamma}_k$ with respect to $\hat{\psi}$, evaluated at $\hat{\psi}$, we define

$$\begin{aligned} g_{3,ik}(\hat{\psi}) &= [\gamma_k'(\hat{\psi})]^2(\hat{\sigma}^2_{a,k}+\hat{\psi})\hat{p}_{T,ik}(1-\hat{p}_{T,ik})\hat{V}(\hat{\psi}) \qquad (37)\\ &= \hat{p}_{T,ik}(1-\hat{p}_{T,ik})\frac{\hat{\sigma}^4_{a,k}}{(\hat{\sigma}^2_{a,k}+\hat{\psi})^3}\hat{V}(\hat{\psi}) \end{aligned}$$

to estimate the effect of estimation of ψ on the MSE. If the estimator of ψ is the pooled estimator, $\hat{\psi}_d$ defined in (19), then $\hat{V}(\hat{\psi})$ in (37) is replaced by $\hat{V}_d$ defined in (36). To estimate the bias of $g_{1,ik}(\hat{\psi})$ for $g_{1,ik}(\psi)$, the corresponding quantity constructed with the true ψ instead of the EGLS estimate, we use

$$\hat{E}[g_{1,ik}(\hat{\psi}) - g_{1,ik}(\psi)] = 0.5g''_{1,ik}(\hat{\psi})\hat{V}(\hat{\psi}) = -g_{3,ik}(\hat{\psi}), \tag{38}$$

where $g''_{1,ik}(\hat{\psi})$ is the second partial derivative of $g_{1,ik}(\hat{\psi})$ with respect to $\hat{\psi}$ evaluated at $\hat{\psi}$. Assembling the components in (34), (37), and (38), we obtain

$$M\hat{S}E_{2,ik} \quad = \quad M\hat{S}E_{1,ik} + 2g_{3,ik}(\hat{\psi}) \tag{39}$$

as the estimator of the MSE of the predictor of the proportion for three digit code i and province k. We define an estimator of the MSE for totals through an argument analogous to the derivation used for proportions. Details about the MSE estimator for totals are omitted.

The estimator of the MSE defined above neglects the effect of the final raking operation. One justification of the Taylor approximation for the raked predictors is that the variance of the small area effect, u_{ik}, converges to zero. Because we do not make this assumption, we do not use the linear approximation to the final raking operation to derive an alternative estimator of the MSE.

6. 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 show that the simulation model satisfies the assumptions of model (1) and specify estimators of the variances. Third, we describe the results of the Monte Carlo experiment.

6.1 Hierarchical Model for Simulation

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 4×10 table of totals from the 2006 Census and estimated province totals from the 2008 LFS were used to define the parameters. Let $\{(\alpha\beta)_{ik}^{cen} : i = 1, \ldots, m; k = 1, \ldots, K\}$ denote maximum likelihood estimates of the interactions in the Census table under the assumption that the Census totals are independent Poisson random variables. The interactions corresponding to three digit code 1 or province 1 are set to zero. Next, fix a parameter vector $\boldsymbol{\lambda}_o = (\alpha_{o,2}, \ldots, \alpha_{o,m}, \beta_{o,1}, \ldots, \beta_{o,K}, \theta_o)'$. Finally, set $T_{ik} = \exp(\alpha_{o,i} + \beta_{o,k} + \theta_o(\alpha\beta)_{ik}^{cen})$, $T_{i.} = \sum_{i=1}^{K} T_{ik}$, $T_{.k} = \sum_{i=1}^{m} T_{ik}$, and $p_{ik} = T_{.k}^{-1} T_{ik}$. The resulting expected values of the proportions and marginal totals are in Table 1.

To induce sampling variability in the direct estimators of the marginal totals, we generate the sample size, $\tilde{n}_k$, in the two digit code in 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 sizes to produce the coefficients of variation of the direct estimators of the province two digit totals in the last row of Table 1. The expected sample sizes are on the horizontal axes of Figures 1-3. The provinces in Table 1 and in Figures 1-3 are listed in increasing order by expected sample size.

Table 1: Expected Values and Coefficients of Variation

Province	PEI	NF	NB	NS	SK	MB	AB	BC	QC	ON	$T_{i.}$
$i = 1$	0.401	0.349	0.340	0.358	0.350	0.348	0.338	0.332	0.382	0.336	99611
$i = 2$	0.194	0.190	0.198	0.206	0.183	0.185	0.212	0.199	0.216	0.222	60438
$i = 3$	0.243	0.253	0.261	0.255	0.256	0.271	0.262	0.281	0.257	0.273	76033
$i = 4$	0.163	0.208	0.201	0.182	0.210	0.195	0.188	0.187	0.144	0.169	48600
$T_{.k}$	1398	4332	6228	8311	8702	10335	29720	35599	73515	106541	284682
CV of $\hat{M}_{.k}$	0.22	0.20	0.20	0.18	0.14	0.12	0.12	0.12	0.11	0.06	

To specify distributions for generating $\boldsymbol{p}_k^*$, we fix $\psi = 0.01726$, the median of the inverses of the expected sample sizes. Then, we generate the vector of true proportions $\boldsymbol{p}_k^*$ from a Dirichlet distribution with probability density function

$$P(x_{1k}, \ldots, x_{mk}) = \left[\frac{\prod_{i=1}^{m} \Gamma(\omega_{ik})}{\Gamma(\omega_o)} \right]^{-1} \prod_{i=1}^{m} x_{ik}^{\omega_{ik}-1}, \tag{40}$$

where $\sum_{i=1}^{m} x_{ik} = 1$, and $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$. To generate vectors of proportions with the desired first and second moments ($E[\boldsymbol{p}_k^*] = \boldsymbol{p}_k$, and $V\{\boldsymbol{p}_k^*\} = \boldsymbol{\Gamma}_{uu,k}\psi$), we set $\omega_o = \psi^{-1} - 1 = 0.01726^{-1} - 1$, and $\omega_{ik} = p_{ik}\omega_o$.

Given $\tilde{n}_k$ and $\boldsymbol{p}_k^*$, the direct estimator of the total in three digit code i and province k depends on a multinomial random vector, $(\tilde{M}_{1k}^{(d)}, \ldots, \tilde{M}_{mk}^{(d)})'$, with probability mass function, $P(\tilde{M}_{1k}^{(d)} = x_{1k}, \ldots, \tilde{M}_{mk}^{(d)} = x_{mk}) = \tilde{n}_k!(\prod_{i=1}^{m} x_{ik}!)^{-1} \prod_{i=1}^{m} (p_{ik}^*)^{x_{ik}}$, where $x_{1k}, \ldots, x_{mk}$ are non-negative integers that sum to $\tilde{n}_k$. The direct estimator of the total employed in three digit code i and province k is $\hat{M}_{ik} = \tilde{n}_k^{-1} \hat{M}_{.k} \tilde{M}_{ik}^{(d)}$. The corresponding direct estimator of the proportion is $\hat{p}_{ik} = \hat{M}_{.k}^{-1} \hat{M}_{ik} = \tilde{n}_k^{-1} \tilde{M}_{ik}^{(d)}$. We explain how this data generating model satisfies the assumptions of model (1) in the next section.

6.2 Second Moments and Variance Estimators

To express the model for simulation in the form of model (1), let $\boldsymbol{e}_k = \hat{\boldsymbol{p}}_k - \boldsymbol{p}_k^*$, and let $\boldsymbol{u}_k = \boldsymbol{p}^* - \boldsymbol{p}_k$. By properties of the multinomial and Dirichlet distributions, $\boldsymbol{e}_k$ and $\boldsymbol{u}_k$ have zero means and are uncorrelated. The variance of $\boldsymbol{u}_k$ is $\boldsymbol{\Gamma}_{uu,k}\psi$, and the variance of $\boldsymbol{e}_k$ is $E\{\tilde{n}_k^{-1}\}\boldsymbol{\Gamma}_{uu,k}(1-\psi)$.

We define estimators of the variances of $\boldsymbol{e}_k$ and $\boldsymbol{u}_k$ to approximate the procedures used in the LFS. The estimator of the sampling variance in the simulation is $\hat{\boldsymbol{\Sigma}}_{ee,k} = \tilde{n}_k^{-1}[\text{diag}(\hat{\boldsymbol{p}}_{T,k}) - \hat{\boldsymbol{p}}_{T,k}\hat{\boldsymbol{p}}'_{T,k}]$. In the LFS, the estimator of ψ is smoothed across several time points, $t = 1, \ldots, d$. In the simulation, we used the procedure described in section 5.2 to combine estimators of ψ from four independently generated data sets.

6.3 Results

Figure 1 shows the ratios of the Monte Carlo MSE's of the predictors, defined in (28), and the GLSM estimators, defined in (10), to the Monte Carlo MSE's of the direct estimators of the proportions. Four ratios are plotted for each province. The expected province sample sample sizes are listed in increasing order on the horizontal axis. The Monte Carlo sample size is 10,000. In the four provinces with the four smallest expected sample sizes, the empirical MSE's of the predictors are approximately $34\% - 51\%$ of the empirical MSE's of the direct estimators (Monte Carlo standard errors of the ratios are 0.006-0.010), while the empirical MSE's of the GLSM estimators are approximately $52\% - 80\%$ of the empirical MSE's of the direct estimators (Monte Carlo SE, 0.010-0.016). In the next four provinces (SK-BC), the empirical MSE's of the predictors are between 57% and 68% of the empirical MSE's of the direct estimators (Monte Carlo SE, 0.008-0.011), while the empirical MSE's of the GLSM estimators are between 116% and 161% of the empirical MSE's of the direct estimators (Monte Carlo SE, 0.024-0.034). In QC, the province with the second largest expected sample size, the MSE's of the predictors are $66\% - 80\%$ of the MSE's of the direct estimators (Monte Carlo SE, 0.005-0.007), and the MSE's of the GLSM estimators are $85\% - 111\%$ of the MSE's of the direct estimators (Monte Carlo SE, 0.014-0.022). In ON, the province with the largest sample size, the empirical MSE's of the predictors are approximately $88\% - 90\%$ of the empirical MSE's of the direct estimators (Monte Carlo SE, 0.007-0.008), while the bias of the GLSM estimator leads to MSE's between 228% and 274% of the MSE's of the direct estimators (Monte Carlo SE, 0.043-0.055). The MSE's of the predictors are uniformly smaller than the MSE's of the direct estimators.

Figure 2 shows ratios of Monte Carlo MSE's of the predictors to Monte Carlo MSE's of the direct estimators for totals and proportions. The relative MSE's for both proportions and totals tend to increase as the expected sample sizes increase. In the smallest province (PEI), the reduction in the MSE is smallest for the cell with the largest expected value. The largest of the expected proportions in the smallest province is 0.401 (Table 1), and the relative MSE for the associated total is 72%.

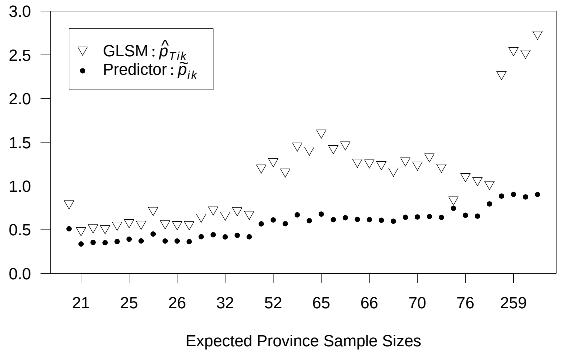

Figure 1: Ratios of Monte Carlo MSE's of predictors and GLSM estimators to Monte Carlo MSE's of direct estimators of proportions. (Monte Carlo sample size: 10,000)

Figure 2 also shows that the reduction in the MSE 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 for totals and proportions. In particular, variability in the direct estimators of the province margins limits the possible reduction in the MSE of totals.

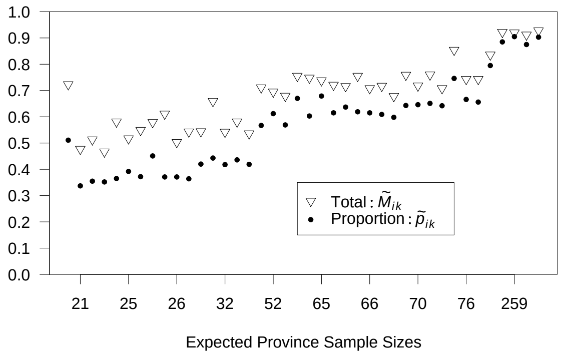

Figure 2: Ratios of Monte Carlo MSE's of predictors to Monte Carlo MSE's of direct estimators. (Monte Carlo sample size: 10,000)

Figure 3 shows the empirical coverages of nominal 95% prediction intervals. The empirical coverages for proportions remain between 94.4% and 96%. Empirical coverages for totals are between 93% and 94% in the smaller provinces and are closer to 95% in the larger provinces. For both totals and proportions, the empirical coverages tend to increase with the expected sample sizes. We do not have an explanation for the increasing trend.

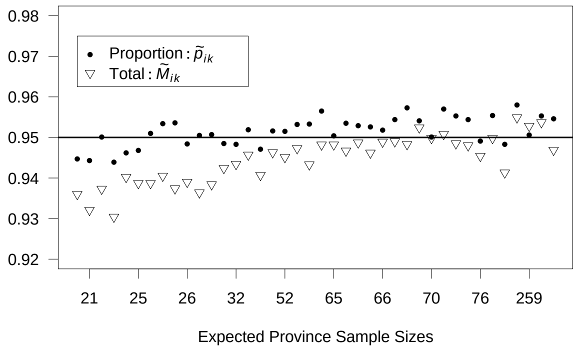

Figure 3: Empirical coverages of nominal 95% prediction intervals. (Monte Carlo sample size: 10,000)

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.

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.

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.

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

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.