Small-Area Estimation: Theory and Practice

Michael Hidiroglou
Statistical Innovation and Research Division, Statistics Canada, 16 th Floor Section D, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada

Abstract

Small area estimation (SAE) was first studied at Statistics Canada in the seventies. Small area estimates have been produced using administrative files or surveys enhanced with administrative auxiliary data since the early eighties. In this paper we provide a summary of existing procedures for producing official small-area estimates at Statistics Canada, as well as a summary of the ongoing research. The use of these techniques is provided for a number of applications at Statistics Canada that include: the estimation of health statistics; the estimation of average weekly earnings; the estimation of under-coverage in the census; and the estimation of unemployment rates. We also highlight problems for producing small-area estimates for business surveys.

KEY WORDS: Small Area, Official Statistics, Fay-Herriot

1. Introduction

Small domain or area refers to a population for which reliable statistics of interest cannot be produced due to certain limitations of the available data. Examples of domains include a geographical region (e.g. a province, county, municipality, etc.), a demographic group (e.g. age x sex), a demographic group within a geographic region**.** The demand for such data small areas has greatly increased during the past few years (Brackstone, 1987). This increase is due to the usefulness of these data in government policy and program development, allocation of various funds and regional planning.

A number of national and regional statistical agencies, including Statistics Canada, have introduced programs aimed at producing estimates for small areas to meet the new demand. Available data to produce such estimates are based on surveys that are not designed for these levels. However, if administrative sources have data at the small area level, and that they are well correlated with variables of interest at the corresponding level, several procedures are available to estimates various parameters of interest for these lower levels**.**

This paper draws on the small area methodology discussed in Rao (2003) and illustrates how some of the estimators have been used in practice on a number of surveys at Statistics Canada. It is structured as follows. Section 2 provides a summary of the primary uses of small area estimates as criteria for computing them. Section 3 defines the notation, provides a number of typical direct estimators, and indirect estimators used in small area estimation. Section 4 provides four examples that reflect the diverse uses of small area estimations at Statistics Canada

2. Primary uses and Criteria for SAE Production

One of the primary objectives for producing small area estimates is provide summary statistics to central or local governments so that they can plan for immediate or future resource allocation. Typical small area estimates include Employment indicators (employed and unemployed), Health indicators (drug use, alcohol use) and Business indicators such as average salary.

The production of small area estimates depends on a number of factors. What the demand for such statistics? What is the commitment and will of the agency to support methodological, systems, and subject matter staff. How much methodology and subject matter expertise exist within the agency. How well correlated are existing auxiliary data with the variables of interest? Is the survey sample size large enough to allow reliable estimates by using both the survey data and the existing auxiliary data? How much bias are the agency and clients willing to tolerate with the estimates,: what are the consequences for making incorrect decisions? The size of the small areas in terms of the number of the units that belong to them is also an important consideration. Small areas that are too small may results in confidentiality breeches. Furthermore, small area estimates may be quite different from statistics based on local knowledge.

3. SAE Estimators

3.1 Introduction

A survey population U consists of N distinct elements (or ultimate units) identified through the labels $j = 1, \ldots, N$. A sample s is selected from U with probability

p(s), and the probability of including the *j-th* element in the sample is π_j . The design weight for each selected unit $j \in s$ is defined as $w_j = 1/\pi_j$. Suppose U_i denotes a domain (or subpopulation) of interest. Denote as $s_i = s \cap U_i$ the part of the sample s that falls in domain U_i .The realized sample size of s_i is a random variable n_i , where $0 \le n_i \le N_i$. Auxiliary data $\boldsymbol{x}$ will either be known at the element level $\boldsymbol{x}_j$ for $j \in s$ or for each small area i as totals $\boldsymbol{X}_i = \sum_{j \in U_i} \boldsymbol{x}_j$ or means $\bar{\boldsymbol{X}}_i = \boldsymbol{X}_i / N_i$.

The problem is to estimate the domain total $Y_i = \sum_{j \in U_i} y_j$ or the domain mean $\bar{Y}_i = Y_i / N_i$, where N_i , the number of elements in U_i may or may not be known. We define y_{ij} to be y_j if $j \in U_i$, and 0 otherwise. An indicator variable a_{ij} is similarly defined: it is equal to one if $j \in U_i$ and 0 otherwise. Note that Y_i can be written as $Y_i = \sum_{j \in U} y_{ij} = \sum_{j \in U} y_j a_{ij}$.

Small area estimation is categorized into two types of estimators: *direct* and *indirect* estimators. A *direct* estimator is one that uses values of the variable of interest, *y*, only from the sample units in the domain of interest. However, a major disadvantage of such estimators is that unacceptably large standard errors may result: this is especially true if the sample size within the domain is small or nil. An *indirect estimator* uses values of the variable of interest from a domain and/or time period other than the domain and time period of interest. Three types of indirect estimators can be identified. A *domain indirect estimator* uses values of the variable of interest from another domain but not from another time period. A *time indirect estimator* uses values of the variable of interest from another time period but not from another domain. An estimator that is both *domain and time indirect* uses values of the variable of interest from another domain and another time period.

An alternative is to use estimators that borrow strength across small areas, by modeling dependent on independent variables across a number of small areas: they are called indirect estimators. Indirect estimators will be quite good (i.e.: indirectly increase the effective sample size and thus decrease the standard error) if the models obtained across small areas still hold at the small area level. Departures from the model will result in unknown biases. There is a wide variety of indirect estimators available, and a good summary is provided in Rao (2003). We will confine ourselves to just a few of them that include the synthetic estimator, and the more well-known composite estimators.

3. 2 Direct Estimation

Let w_j be the design weight associated with $j \in s$. The Horvitz-Thompson is the simplest direct estimator. If the small area total Y_i is to be estimated for small area U_i , then the corresponding Horvitz-Thompson estimator is given by $\hat{Y}_{i,HT} = \sum_{j \in s_i} w_j y_j$ provided that the realized sample size n_i is non-zero.

Auxiliary information can be available either at the population level or at the domain level. If it available at the population level, then we used the Generalized Regression Estimator (GREG) given by

$\hat{Y}_{i,GR} = \boldsymbol{X}' \tilde{\boldsymbol{\beta}}_{i,GREG} + \left(\hat{Y}_{i,HT} - \hat{\boldsymbol{X}}'_{HT} \tilde{\boldsymbol{\beta}}_{i,GREG} \right)$ where

$\boldsymbol{X}' = \sum_{i=1}^{m} \sum_{j \in U_i} \boldsymbol{x}_j$, $\hat{\boldsymbol{X}}'_{HT} = \sum_s \boldsymbol{x}'_k / \pi_k$, and

$\tilde{\boldsymbol{\beta}}_{i,GREG}$ is the set of regression coefficient obtained by regressing y_{ij} on $\boldsymbol{x}_j$. That is

$$\tilde{\boldsymbol{\beta}}_{i,GREG} = \left(\sum_s \frac{w_j \boldsymbol{x}_j \boldsymbol{x}'_j}{c_j} \right)^{-1} \sum_s \frac{w_j \boldsymbol{x}_j y_{ij}}{c_j},$$

where c_j is a specified constant (c_j>0).

The straight GREG is estimator is not efficient, and it is better to use regression estimators that use auxiliary data available as close possible to the small areas of interest. One such estimator is the domain–specific GREG that uses auxiliary data at the domain level. It is given by $Y^*_{i,GR} = \boldsymbol{X}'_i \hat{\boldsymbol{\beta}}_{i,GREG} + \left(\hat{Y}_{i,HT} - \hat{X}'_{i,HT} \hat{\boldsymbol{\beta}}_{i,GREG} \right)$

where $\hat{\boldsymbol{\beta}}_{i,GREG} = \left(\sum_{s_i} w_j \boldsymbol{x}_j \boldsymbol{x}'_j / c_j \right)^{-1} \sum_{s_i} w_j \boldsymbol{x}_j y_j / c_j$.

An estimator that is approximately p-unbiased as the overall sample size increases but uses *y*-values outside the domain is the modified direct estimator given by $\hat{Y}_{i,SR} = \boldsymbol{X}'_i \hat{\boldsymbol{\beta}}_{GREG} + \left(\hat{Y}_{i,HT} - \hat{X}'_{i,HT} \hat{\boldsymbol{\beta}}_{GREG} \right)$ where

$\hat{\boldsymbol{\beta}}_{GREG} = \left(\sum_s w_j \boldsymbol{x}_j \boldsymbol{x}'_j / c_j \right)^{-1} \sum_s w_j \boldsymbol{x}_j y_j / c_j$ .This estimator is also referred to in Woodruff (1966), and Battese, Harter, and Fuller (1988) as the "survey regression estimator".

Hidiroglou and Patak (2004) compared a number of the direct estimators. One of their conclusions was that the direct estimators would be best if the domains of interest coincided as closely as possible with the design strata.

3.2 Indirect Estimation

Some of the most widely used indirect estimators have been the synthetic estimator, the regression-adjusted synthetic, the composite estimator, and the sample-dependent estimator.

The synthetic estimator uses reliable information of a direct estimator for a large area that spans several small areas, and this information is used to obtain an indirect estimator for a small area. It is assumed that the small areas have the same characteristics as the large area: Gonzalez (1978) provides a good account how these estimators were obtained, and used to obtain unemployment statistics at levels lower than those planned in the survey design. The National Center for Health Statistics (1968) in the United States pioneered the use of synthetic estimation for developing state estimates of disability and other health characteristics from the National Health Interview Survey (NHIS). Sample sizes in most states were too small to provide reliable direct state estimates.

Levy (1971) used mortality data to compute average relative errors of synthetic estimates for States. He used the regression-adjusted synthetic estimator to account for local variation by combining area-specific covariates with the synthetic estimator. These covariates attempted to attenuate the magnitude of potential relative bias associated with the synthetic estimator.

The potential bias associated with indirect estimators $\hat{Y}_{i,INDIR}$ can be attenuated by combining them with the direct estimators $\hat{Y}_{i,DIR}$ via a weighted average. The resulting combined estimator is given by

$$\hat{Y}_{i,COMB} = \phi_i \hat{Y}_{i,DIR} + (1-\phi_i)\hat{Y}_{i,INDIR}$$

where ϕ_i $\left(0 \le \phi_i \le 1\right)$. The optimal ϕ_i^* is determined by minimizing the MSE of $\hat{Y}_{i,COMB}$. The resulting composite estimator has a mean square error which is smaller than that of either component estimator. Schaible (1978) noted that the composite estimator is insensitive to poor estimates of the optimum weight. This insensitivity depends on the relative sizes of the mean square errors of the component estimators. The composite estimator is most insensitive when the mean square errors of the two component estimators do not differ greatly. Simple weighting factors for the composite estimators that depend on the realized domain size were given by Drew, Singh and Choudhry (1982), and by Hidiroglou and Särndal (1985)

Small area estimators are split into two main types, depending on how models are applied to the data within the small areas: these two types are known as *area level* and *unit level*. Small area estimators are based on area level computations if models link small area means of interest (*y*) to area-specific auxiliary variables (such as $\boldsymbol{x}$ sample means). They are based on unit level computations if the models link unit values of interest to unit-specific auxiliary variables. Area based small area estimators are computed if the unit level area data are not available. They can also be computed if the unit level data are available by summarizing them at the appropriate area level.

3.2.1 Area Model

One of the most widely used area based level small area estimator was given by Fay and Herriot (1979) small. Population totals ($Y_i = \sum_{j \in U_i} y_j$) or means ($\bar{Y}_i = Y_i / N_i$), where N_i is the number of elements in small area U_i , can be estimated. The Fay-Herriot methodology is usually presented as an estimator of the small area population mean $\bar{Y}_i$ for a given small area U_i where $i = 1, \ldots, m$. The Fay-Herriot estimator for small area U_i is a linear combination of a direct estimator (say $\hat{\bar{Y}}_{i,DIR}$) and a synthetic estimator (say $\hat{\bar{Y}}_{i,SYN}$). The direct estimator of the population mean $\bar{Y}_i$ is given by $\hat{\bar{Y}}_{i,DIR} = \hat{Y}_{i,DIR} / \hat{N}_{i,DIR}$ where $\hat{Y}_{i,DIR} = \sum_{j \in s_i} \tilde{w}_j y_j$ and $\hat{N}_{i,DIR} = \sum_{j \in s_i} \tilde{w}_j$. The weight $\tilde{w}_j$ associated with the *j-th* unit can be the design weight w_j (i.e. $\tilde{w}_j = w_j$) or a final weight that reflects any adjustment (i.e.: non-response, calibration, or a product thereof) made to the design weight.

The synthetic portion is estimated as the product of a given auxiliary population mean row-vector (say $\bar{\boldsymbol{Z}}_i' = \sum_{j \in U_i} \boldsymbol{z}_j' / N_i$) for the *i-th* small area of interest times an estimated regression vector (say $\hat{\bar{\boldsymbol{\beta}}}_{FH}$, where *FH* stands for Fay-Herriot). The auxiliary data row-vector $\boldsymbol{z}_j'$ is known for all units in the population

small area U_i. The regression vector $\hat{\tilde{\boldsymbol{\beta}}}_{FH}$ is computed across a number of small areas in such a way that the model linking the variable of interest (the mean $\hat{\bar{Y}}_{i,DIR}$) auxiliary data also holds at the small area level. The Fay-Herriot estimator of a given population mean $\bar{Y}_i$ is estimated as:

$$\hat{\tilde{\bar{Y}}}_{i,FH} = \gamma_i \hat{\bar{Y}}_{i,DIR} + (1-\gamma_i)\bar{\boldsymbol{Z}}_i'\hat{\tilde{\boldsymbol{\beta}}}_{FH} \qquad (3.1)$$

The two components (direct estimator and synthetic estimator) of (3.1) are weighted γ_i and $(1-\gamma_i)$ where $\gamma_i = \sigma_v^2/(\sigma_v^2+\psi_i)$. The regression vector $\hat{\tilde{\boldsymbol{\beta}}}_{FH}$ and γ_i depend on the population variance $\bar{\psi}_{i,DIR}$ of the direct estimator $\hat{\bar{Y}}_{i,DIR}$ and the model variance σ_v^2. Although the sampling variance of $\hat{\bar{Y}}_{i,DIR}$ is easy to compute, it may be unstable if the domain sizes are small. This is repaired with a smoothing of the estimated variances. We denote the smoothed variances as $\hat{\bar{\psi}}_{i,DIR}$. The estimated model variance $\hat{\sigma}_v^2$ and $\hat{\tilde{\boldsymbol{\beta}}}_{FH}$ are computed recursively. Details of the required computations for obtaining $\hat{\sigma}_v^2$ can be found in of Rao (2003, pp. 118-119). The estimated regression vector $\hat{\tilde{\boldsymbol{\beta}}}_{FH}$ and the factor $\hat{\tilde{\gamma}}_i$ are given by:

$$\hat{\tilde{\boldsymbol{\beta}}}_{FH} = \left[\sum_{i=1}^{D} \frac{\bar{\boldsymbol{Z}}_i'\bar{\boldsymbol{Z}}_i}{\hat{\bar{\psi}}_{i,DIR}+\hat{\sigma}_v^2}\right]^{-1}\left[\sum_{i=1}^{D} \frac{\bar{\boldsymbol{Z}}_i'\hat{\bar{Y}}_{i,DIR}}{\hat{\bar{\psi}}_{i,DIR}+\hat{\sigma}_{v\ i}^2}\right] \qquad (3.2)$$

and

$$\hat{\tilde{\gamma}}_i = \hat{\sigma}_v^2/\left(\hat{\bar{\psi}}_{i,DIR}+\hat{\sigma}_v^2\right) \qquad (3.3)$$

respectively.

The Fay-Herriot estimator $\hat{\bar{Y}}_{i,FH}$ can also be expressed as:

$$\hat{\bar{Y}}_{i,FH} = \bar{\boldsymbol{Z}}_i'\hat{\tilde{\boldsymbol{\beta}}}_{FH} + \hat{\tilde{\gamma}}_i\left(\hat{\bar{Y}}_{i,DIR} - \bar{\boldsymbol{Z}}_i'\hat{\tilde{\boldsymbol{\beta}}}_{FH}{}^{FH}\right). \qquad (3.4)$$

This form of the Fay-Herriot estimator is very similar to the "normal" regression estimator

$$\hat{\bar{Y}}_{i,REG} = \bar{\boldsymbol{Z}}_i'\hat{\tilde{\boldsymbol{\beta}}}_{REG} + \left(\hat{\bar{Y}}_{i,EXP} - \bar{\boldsymbol{Z}}_i'\hat{\tilde{\boldsymbol{\beta}}}_{REG}\right) \qquad (3.5)$$

given in Cochran (1977), where the estimated regression vector is given by

$$\hat{\tilde{\boldsymbol{\beta}}}_{REG} = \left(\sum_{i=1}^{D}\bar{\boldsymbol{Z}}_i'\bar{\boldsymbol{Z}}_i/\hat{\bar{\psi}}_{i,EXP}\right)^{-1}\left(\sum_{i=1}^{D}\bar{\boldsymbol{Z}}_i'\ \hat{\bar{\psi}}_i^{EXP}/\hat{\bar{\psi}}_{i,EXP}\right)$$

and $\hat{\bar{\psi}}_i^{EXP} = \sum_{j\in U_i} w_j y_j \Big/ \sum_{j\in U_i} w_j$ is the simple estimator of the mean involving the design weights w_j. The computations required to obtain the normal regression estimator do not involve estimating any variance components.

3.2.2 Unit Model

The unit model originates with Battese, Harter and Fuller (1988). They used the nested error regression model to estimate county crop areas using sample survey data in conjunction with satellite information. Their model is given by

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij} \qquad (3.5)$$

where they assumed that $v_i \overset{iid}{\sim}(0,\sigma_v^2)$; $e_{ij} \overset{iid}{\sim}(0,\sigma_e^2)$, *i=1,…,m* and $j=1,\ldots,n_i$. The small areas of interest in Battese, Harter and Fuller (1988) were 12 counties (*m=12*) in North-Central Iowa. Each county was divided into area segments and the areas under corn and soybeans were ascertained for a sample of segments by interviewing farm operators. The number of sampled segments in a county n_i ranged from 1 to 6. Auxiliary data were in the form of numbers of pixels (a term used for "picture elements" of about 0.45 hectares) classified as corn and soybeans were also obtained for all the area segments, including the sampled segments, in each county using LANDSAT satellite readings.

The resulting sample mean using (3.5) is given by

$$\bar{y}_{i.} = \bar{\boldsymbol{x}}_{i.}'\boldsymbol{\beta} + v_i + \bar{e}_{i.} \qquad (3.6)$$

where $\bar{y}_{i.}, \bar{\boldsymbol{x}}_{i.}'$ and $\bar{e}_{i.}$ are the means of the associated n_i (*y,* $\boldsymbol{x}$**)** observations and *e*-residuals. Battese et al. (1988)'s objective was to estimate the conditional population mean given the realized cluster (county) effect. Under the assumption of model (3.5), the conditional population mean is given by

$$\bar{Y}_{i.} = \bar{\boldsymbol{X}}_{i.}'\boldsymbol{\beta} + v_i \qquad (3.7)$$

where $\bar{Y}_{i.}, \bar{\boldsymbol{X}}_{i.}'$ are the population means of the associated N_i observations ($y_{ij}, \boldsymbol{x}_{ij}$)in the *i-th* sampled cluster U_i. The corresponding predictor $\tilde{y}_{i.}$ for the county mean crop area per segment is $\bar{\boldsymbol{X}}_{i.}'\tilde{\boldsymbol{\beta}} + \tilde{v}_{i.}$

where $\tilde{v}_{i.} = n_i^{-1}\sum_{j=1}^{n_i}\left(y_{ij} - \boldsymbol{x}_{ij}'\tilde{\boldsymbol{\beta}}\right)\gamma_i$ with

$$\tilde{\boldsymbol{\beta}}_{BHF} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}' - \gamma_i\bar{\boldsymbol{x}}_{i.}\bar{\boldsymbol{x}}_{i.}'\right)\right)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\boldsymbol{x}_{ij}y_{ij} - \gamma_i\bar{\boldsymbol{x}}_{i.}\bar{y}_{i.}\right) \qquad (3.8)$$

and $\gamma_i = \sigma_v^2 \left(\sigma_v^2 + n_i^{-1}\sigma_e^2\right)^{-1}$.

The resulting best linear unbiased prediction (BLUP) estimator is $\tilde{y}_{i,BHF} = \gamma_i \bar{y}_{i.} + \left(\bar{\boldsymbol{X}}_{i.}' - \gamma_i \bar{\boldsymbol{x}}_{i.}'\right)\tilde{\boldsymbol{\beta}}_{BHF}$ for the *i-th* small area. However, the variance components σ_v^2 and σ_e^2 are not known. Battese et al (1988) use the well-known-method of fitting-of-constants to estimate them. The resulting estimator of the *i-th* area sample mean is known as the EBLUP estimator, because the variance components were estimated.

Prasad and Rao (1990) derived an approximation to $o(m^{-1})$ for the model based mean squared error of the Battese-Harter-Fuller estimator, and also obtained its estimator to $o(m^{-1})$ as well. Prasad-Rao (1999) were the first to include the survey weights in the unit level model: they labelled their estimator as a pseudo-EBLUP estimator of the small area mean $\bar{Y}_i$. The Prasad-Rao estimator of $\bar{Y}_i$ is given by

$$\tilde{\bar{Y}}_{i,PR} = \bar{\boldsymbol{X}}_i'\tilde{\boldsymbol{\beta}}_{PR} + \gamma_{iw}\left(\bar{y}_{iw} - \bar{\boldsymbol{x}}_{iw}'\tilde{\boldsymbol{\beta}}_{PR}\right) \qquad (3.9)$$

where $\gamma_{iw} = \sigma_v^2 / \left(\sigma_v^2 + \sigma_e^2 \sum_{j\in s_i} \tilde{w}_j^2\right)$ with $\bar{y}_{iw} = \sum_{j\in s_i} \tilde{w}_j y_j$; $\tilde{w}_{ij} = w_{ij}^* / \sum_{j\in s_i} w_{ij}^*$ and w_{ij}^* are calibrated weights, and $\tilde{\boldsymbol{\beta}}_{PR}$ is given by

$$\tilde{\boldsymbol{\beta}}_{PR} = \left(\sum_{i=1}^{m} \gamma_{iw} \bar{\boldsymbol{x}}_{iw} \bar{\boldsymbol{x}}_{iw}'\right)^{-1} \sum_{i=1}^{m} \gamma_{iw} \bar{\boldsymbol{x}}_{iw} \bar{y}_{iw} \qquad (3.10)$$

Prasad and Rao (1999) also provided model based expressions for the MSE of their estimator when it included the estimated variance components σ_v^2 and σ_e^2.

The sum of small area estimates do not necessarily add up to the corresponding direct estimator. You-and Rao (2002) proposed an estimator of $\boldsymbol{\beta}$ that ensures self-benchmarking of the small area estimates to the corresponding direct estimator. Their estimator is given by

$$\bar{Y}_{i,YR} = \bar{\boldsymbol{X}}_i'\hat{\boldsymbol{\beta}}_{YR} + \gamma_{iw}\left(\bar{y}_{iw} - \bar{\boldsymbol{x}}_{iw}'\hat{\boldsymbol{\beta}}_{YR}\right) \qquad (3.10)$$

where

$$\tilde{\boldsymbol{\beta}}_{YR} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n_i} \tilde{w}_{ij}\boldsymbol{x}_{ij}\left(\boldsymbol{x}_{ij} - \gamma_{iw}\bar{\boldsymbol{x}}_{iw}\right)'\right)^{-1} \sum_{i=1}^{m}\sum_{j=1}^{n_i} \tilde{w}_{ij}\left(\boldsymbol{x}_{ij} - \gamma_{iw}\bar{\boldsymbol{x}}_{iw}\right)y_{ij}.$$

Replacing σ_v^2 and σ_e^2 by $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$, we obtain a survey-weighted estimator of $\boldsymbol{\beta}$ (say $\tilde{\boldsymbol{\beta}}_{YR}$). The resulting "pseudo-EBLUP" estimator $\hat{\bar{Y}}_{i,PR}$ is given by $\hat{\bar{Y}}_{i,PR} = \bar{\boldsymbol{X}}_i'\hat{\boldsymbol{\beta}}_{PR} + \hat{\gamma}_{iw}\left(\bar{y}_{iw} - \bar{\boldsymbol{x}}_{iw}'\hat{\boldsymbol{\beta}}_{PR}\right)$. Note that the self-benchmarking property means that the sum of the estimated small area totals is equal to the direct estimator of the overall total *Y*. That is,

$$\sum_{i=1}^{m} N_i \hat{\bar{Y}}_{i,PR} = \hat{Y}_w + \left(\boldsymbol{X} - \boldsymbol{X}_w\right)'\hat{\boldsymbol{\beta}}_w$$

where $\hat{Y}_w = \sum_{i=1}^{m} N_i \hat{\bar{Y}}_{i,PR}\hat{Y}_w = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \tilde{w}_{ij}\, y_{ij}$ and $\hat{\boldsymbol{X}}_w$ is similarly defined.

4. Applications

4.1 Canadian Community Health Survey: Area model

The Canadian Community Health Survey CCHS is a cross-sectional health survey carried out by Statistics Canada since 2001.The survey operates on a two-year collection cycle. The first year of the survey cycle "x.1" is a large sample (130,000 persons), general population health survey, designed to provide reliable estimates at the health region (sub-provincial areas defined in terms of Census results), provincial and national levels. This portion of the survey collects information related to health status, health care utilization and health determinants for the Canadian population. The second year of the survey cycle "x.2" has a smaller sample (30,000 persons) and is designed to provide provincial and national level results on specific focused health topics.

The CCHS is based on a multiple frame (two frames) sampling design of that uses. The first one, used as the primary frame, is the area frame designed for the Canadian Labour Force Survey. This survey is basically a two-stage stratified design that uses probability proportional to size without replacement at each stage. Face to face interviews take place with individuals selected from that frame. The second frame uses a list frame of telephone numbers in some of the Health Regions for cost reasons. Individuals selected in that frame are interviewed by telephone.

The area frame uses the *Labour Force Frame.* This resulting sample is a two-stage stratified cluster. Sampling in that frame is carried out in three steps. Firstly, a list of the dwellings that were or had been in

scope to the Labour Force sample is identified. Secondly, a sample of dwellings was selected from this list. The households in the selected dwellings then formed the sample of households. The majority (88%) of the targeted sample was selected from the area frame. Lastly, respondents are randomly selected from households in this frame. Although a single individual is normally randomly selected from each household, the requirement to over sample youths results in a second member of a number of households to be selected as well. Face-to-face interviews are carried out with the selected respondents.

The telephone frame is mainly based on a stratified version (Health Regions) of the Canada Phone directory. Simple random sampling takes place within each of the resulting strata. Random digit dialling is carried out in five HRs and the three Territories.

The direct estimator of a population total Y_i for a given domain i is given by $\hat{Y}_i^{DIR} = \sum_{j \in S_i} \tilde{w}_j^* y_j$ where $\tilde{w}_j^*$ represents the overall weight that incorporates the multiple frame nature of the sampling design, non-response adjustments at each stage, where appropriate, and the calibration (age groups 12 to 19, 20 to 29, 30 to 44, 45 to 64 and 65 or older for each sex within each health region and province). More details of this sampling design are available in Béland (2002).

Estimates of various population parameters can be produced for different domains. In the present example, taken from Hidiroglou, Singh and Hamel (2007), our parameter of interest is the proportion of alcohol abuser within the previously stated domains belonging to the province of British Columbia using the two year (2000-2001) CCHS sample. The associated sample had 18,302 observations with domain sample size ranging from 20 to 238 for the 200 domains. Figure 4.1 provides an idea of how the Health Regions are delineated in British Columbia.

The i-th domain is a cross-classification of health regions r ($r = 1,\ldots, 20$) and age-sex groups a ($a = 1,\ldots, 10$). The direct estimator of proportion of alcohol abuse is given by $\hat{p}_{r,a}^{DIR} = \hat{Y}_{r,a}^{DIR} / \hat{N}_{r,a}^{DIR}$ where $\hat{N}_{r,a}^{DIR} = \sum_{j \in S_{r,a}} \tilde{w}_j^*$. Given that, for domain *ra,*

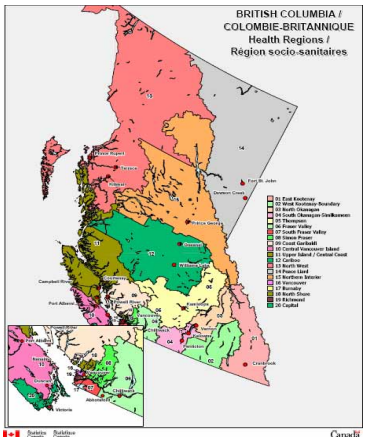

Figure 4.1: Health areas in British Columbia

$\hat{\psi}_{r,a}^{DIR}$ denotes the estimated variance for $\hat{p}_{r,a}^{DIR}$ under the sampling design, the associated estimated design effect is given by $deff_{r,a}^{DIR} = \hat{\psi}_{r,a}^{DIR} / \left(\hat{p}_{r,a}^{DIR} \left(1 - \hat{p}_{r,a}^{DIR} \right) / n_{r,a} \right)$. The smoothed design effect over all *I=200* domains is given by $\overline{def}^{DIR} = \sum_i deff_i^{DIR} / I$. The estimated coefficient of variation, $cv\left(\hat{p}_{r,a}^{DIR} \right)$, for $\hat{p}_i^{DIR}$ for a given domain *i* is $\sqrt{\overline{def}^{DIR} \left(\hat{p}_{r,a}^{DIR} \left(1 - \hat{p}_{r,a}^{DIR} \right) / n_{r,a} \right)} / \hat{p}_{r,a}^{DIR}$.

The common mean model is the simplest one that can be implemented using the Fay-Herriot (1979) methodology. This model assumes that the proportion of alcohol abuse is the same within each of the twenty Health Regions for a given age-sex group: that is, the linking model is given by $P_{r,a} = \beta_a + v_{r,a}$ where $P_{r,a}$ is the unknown population proportion of interest, and β_a is the common mean across the health regions for the a-th age-sex group. The corresponding sampling model is given by $\hat{p}_{r,a}^{DIR} = P_{r,a} + e_{r,a}$. The resulting small area estimate for the *ra*-th domain is given by $\hat{p}_{r,a}^{EBLUP} = \hat{\gamma}_{r,a} \hat{p}_{r,a}^{DIR} + \left(1 - \hat{\gamma}_{r,a} \right) \hat{\beta}_{r,a}$, where $\hat{\gamma}_{r,a} = \dfrac{\hat{\sigma}_v^2}{\tilde{\psi}_{r,a}^{DIR} + \hat{\sigma}_v^2}$ (see Rao 2003, p. 116). The $\tilde{\psi}_{r,a}^{DIR}$ term given by $\tilde{\psi}_{r,a}^{DIR} = \overline{def}^{DIR} \dfrac{\hat{p}_{r,a}^{DIR} \left(1 - \hat{p}_{r,a}^{DIR} \right)}{n_{r,a}}$ is obtained using the smoothed design effect $\overline{def}^{DIR} = \sum_i deff_i^{DIR} / I$ over the *I=200* domains. The $\hat{\sigma}_v^2$ term is obtained from the Fay-Herriot methodology: computational details for estimating $\hat{\sigma}_v^2$ can be found in Rao (2003, p. 118). The estimated coefficient of variation $cv\left(\hat{p}_{r,a}^{EBLUP} \right)$ for

$\hat{p}_{r,a}^{EBLUP}$ is given by $\frac{\sqrt{mse\left(\hat{p}_{r,a}^{EBLUP}\right)}}{\hat{p}_{r,a}^{EBLUP}}$, where $mse\left(\hat{p}_{r,a}^{EBLUP}\right) = \frac{\hat{\sigma}_v^2 \tilde{\psi}_{r,a}^{DIR}}{\tilde{\psi}_{r,a}^{DIR} + \hat{\sigma}_v^2}$ represents the estimated leading term of $MSE\left(\hat{p}_{r,a}^{EBLUP}\right)$. Figure 4.2 is a graph between the estimated coefficients of variation resulting for the direct and indirect estimation

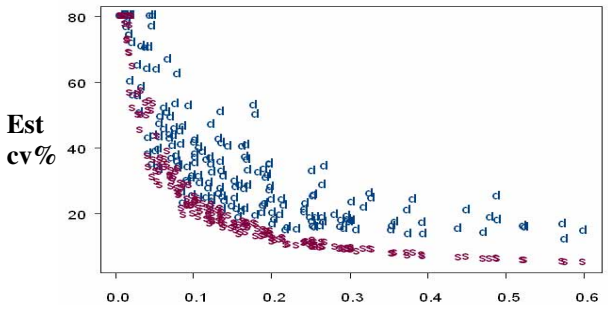

Figure 4.2: Estimated coefficients of variation for the direct (blue) and EBLUP (red) estimators of proportion

4.2 Canadian Survey of Employment Payroll and Hours: Unit model

The Canadian Survey of Employment, Payrolls and Hours (SEPH) collects and publishes on a monthly basis, estimates of payrolls, employment, paid hours and earnings at detailed industrial and geography levels. Estimators for average weekly earnings (AWE) have been produced since the early nineties by SEPH. These estimates have been produced via the generalized regression (GREG) estimator using a combination of survey and payroll deduction (administrative) data provided to Statistics Canada by the Canada Tax department. The GREG estimator is approximately design unbiased (ADU).

SEPH is currently being redesigned to redefine primary domains of interest, as well as incorporate improvements on the use of the administrative data. The resulting sample, estimated to be between 11,00 to 20,00 establishments (depending on budget constraints) will be allocated to the newly defined strata, defined as cross-classifications of geography (provinces) and industry (NAICS3), so that the resulting GREG estimates for AWE satisfy coefficients of variation. The design strata are also referred to model groups since the GREG estimators are computed at these levels as well. Estimates below this level can be obtained using domain estimation. As the sample associated will be relatively small (or non-existent), the reliability associated with the GREG estimators could be unacceptably large measures of error.

Rubin et al. (2007) investigated whether Small Area Estimation (SAE) procedures could be used to produce estimates for AWE with reasonably good estimated mean squared errors for lower levels, namely industry groups at the North American Industry Classification System (NAICS4) level 4 and geography at the level of province, that is, the "NAICS 4 x province" domains. The Average Weekly Earnings for a population domain i (U_i) is given by

$$\bar{Y}_i = \sum_{j \in U_i} E_{ij} y_{ij} / \sum_{j \in U_i} E_{ij}$$

where y_{ij} is the average weekly earnings and E_{ij} is the average number of employees within the j-th establishment within that domain.

A Monte Carlo study was carried out to evaluate the properties of the GREG estimator and a number of SAE estimators. The y-values for the population used for the study were created for twelve months for twelve months representing the January to December 2005 calendar year. In sample y-values were kept as is, and the kept as is and the y-values for the out-of-sample units were synthesized using the nearest neighbour using the average number of employment and average monthly earnings (available for the whole population). Some 100o samples were then independently sampled from each of the twelve generated populations, preserving the longitudinal aspect of SEPH (i.e.: sample rotation of one-twelfth of the sample on a monthly basis). Summary statistics based on the specific estimators, $\bar{y}_{i,EST}^{(r)}$, used of the i-th small area ($i=1,\ldots,I$) computed from the Monte Carlo, included the average relative bias (ARB), $\frac{1}{I}\sum_{i=1}^{I}\left|\frac{1}{R\bar{Y}_i}\sum_{r=1}^{R}\left(\bar{y}_{i,EST}^{(r)} - \bar{Y}_i\right)\right|$, and the average root relative mean square error (ARMSE), $\frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{R\bar{Y}_i}\sum_{r=1}^{R}\left(\bar{y}_{i,EST}^{(r)} - \bar{Y}_i\right)^2\right)^{0.5}$.

Estimators considered in the Rubin et al. (2007) simulation included the GREG, the Prasad-Rao (1999) pseudo-EBLUP unit level, and the You-Rao (2002) pseudo-EBLUP area level SAE estimators given in Section 3.0. The GREG estimator is given by

$$\bar{y}_{i,GREG} = \sum_{U_i} \tilde{E}_{ij} \boldsymbol{x}_{ij}' \hat{\boldsymbol{\beta}} + \sum_{s_i} w_{ij} \tilde{E}_{ij} \left(y_{ij} - \boldsymbol{x}_{ij}' \hat{\boldsymbol{\beta}}\right) \quad (4.1)$$

with $\boldsymbol{x}_{ij}' = \left(1, x_{ij}\right)$. Here x_{ij} is the average monthly earnings associated with the j-th sampled establishment within domain U_i, and $\hat{\boldsymbol{\beta}}$ is the

regression estimator resulting from the model $y_{ij} = \boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}} + e_{ij}$, with $e_{ij} \overset{iid}{\sim} (0, \sigma_e^2 / E_{ij})$.

Figure 3 and 4 provide the ARB and ARMSE respectively for construction domains in Canada for 2005. The GREG estimator has the smallest ARB amongst the three estimators. The Prasad-Rao (1999) is the best estimator in terms of ARMSE. This is reasonable on account that the You-Rao (2002) estimator loses efficiency on account of its benchmarking property.

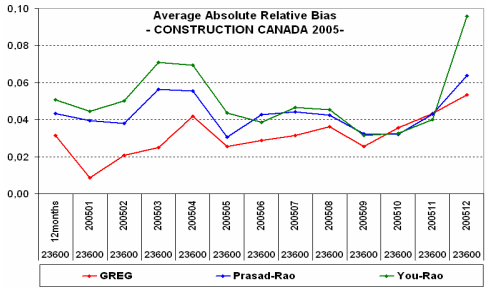

Figure 4.3: Average absolute relative bias for construction domains in Canada for 2005

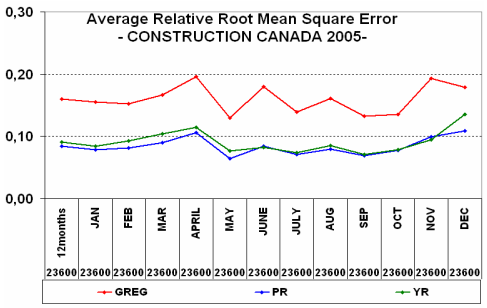

Figure 4.4: Average relative root mean square error for construction domains in Canada for 2005

4.3 Canadian Census of Population under coverage

The Census of Canada is conducted every five years. One objective is to provide the Population Estimates Program with accurate baseline counts of the number of persons by age and sex for specified geographic areas. However, not all persons are correctly enumerated. Two errors that occur are undercoverage - exclusion of eligible persons - and over coverage - erroneous inclusion of persons. This undercoverage varies between 2 and 3 %.

A special survey, known as the Reverse Record Check (RRC), with a sample size of 60,000 persons, estimates the net number of persons missed by the Census. This net number combines two types of coverage errors: the gross number of persons missed by the Census (Undercount) and the gross number of persons erroneously included in the final Census count (Overcount). The sample size of the RRC is designed to produce reliable direct estimates for the provinces (including the two Territories),and eight age - sex groups, with age categories are less than 19, 20 to 29, 30 to 44, and 45 and over at the national level. The cross tabulation of these two marginal tabulations results in *m*= 96 (12*8) cells. These cells are considered as small areas because they have too few observations to sustain reliable direct estimates. The objective is to use small area techniques to improve the reliability of the cell estimates. Dick (1995) applied the Fay-Herriot methodology for this purpose.

For the *i-th* cell (small area), we define the following quantities. The true (but unknown) Census count is denoted as T_i, and the corresponding observed Census count as C_i. This means that the difference $(T_i - C_i)$ is the missed unknown net undercoverage count (M_i). This net undercoverage count is estimated by the RRC for the i-th small area is $\hat{M}_i$. The true count T_i can be expressed as the product of the observed count C_i and the true adjustment factor $\theta_i = (M_i + C_i)/C_i = T_i / C_i$. The true adjustment factor can be estimated directly as $y_i = (\hat{M}_i + C_i)/C_i$. However, the direct estimator $\hat{M}_i$ may not be reliable. The problem is cast into a Fay-Herriot context as follows.

The sampling model can be written as $y_i = \theta_i + e_i$ where we assume that $E_p(e_i) = 0$ and $V_p(e_i) = \psi_i$, where ψ_i is assumed to be known. The linking model is given by $\theta_i = \boldsymbol{z}_i'\boldsymbol{\beta} + v_i$, where $\boldsymbol{z}_i$ is a set of auxiliary variables, and $v_i \overset{iid}{\sim} (0, \sigma_v^2)$.

The resulting Fay-Herriot estimator is given as $\hat{\theta}_{i,FH} = \boldsymbol{z}_i'\hat{\boldsymbol{\beta}}_{FH} + \hat{\gamma}_i\left(y_i - \boldsymbol{z}_i'\hat{\boldsymbol{\beta}}_{FH}\right)$ where $\hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \psi_i)$.

The sampling variances are not known, but can be estimated as from $\hat{\psi}_i = \mathrm{v}(y_i)$ given the sampling plan for the RRC. As these variances are for domains, they will be tend to be variable. Dick (1995) smoothed them

by using $\log\left(\mathrm{v}(\hat{M}_i)\right) = \alpha + \beta \log\left(C_i\right) + \eta_i$ where it is assumed that $\eta_i \overset{iid}{\sim} N\left(0, \zeta^2\right)$.

The smoothed estimate of variance for the *i-th* small area is $\tilde{\mathrm{v}}(\hat{M}_i) = \exp\left(\hat{\alpha} + \hat{\beta} \log\left(C_i\right)\right)$. Hence, the smoothed variance of $y_i = 1 + \hat{M}_i / C_i$ is $\tilde{\psi}_i = \tilde{\mathrm{v}}(\hat{M}_i) / C_i^2$.

Replacing the unknown ψ_i by $\tilde{\psi}_i$ leads to $\tilde{\theta}_{i,FH} = z_i' \tilde{\boldsymbol{\beta}}_{FH} + \tilde{\gamma}_i \left(y_i - z_i' \tilde{\boldsymbol{\beta}}_{FH}\right)$ where $\tilde{\sigma}_v^2$ and $\tilde{\boldsymbol{\beta}}_{FH}$ are solved iteratively using the algorithm given in the appendix.

State which variables used and for Census (2011). For further details see You, Rao and Dick (2002)

The above methodology was used to estimate the 2001 Canadian Census undercoverage. The final *z*-variables used in the linking model (4.3) were Yukon, Nunavet, Male 20 to 29, Male 30 to 44, Female 20 to 29, British Colombia renters, Ontario renters and North West Territories renters.

Figure 1 displays the direct and FH estimates of undercoverage ratios by the domain sample sizes. Figure 2 displays the corresponding coefficients of variation (CV) of the direct and FH HB estimates.

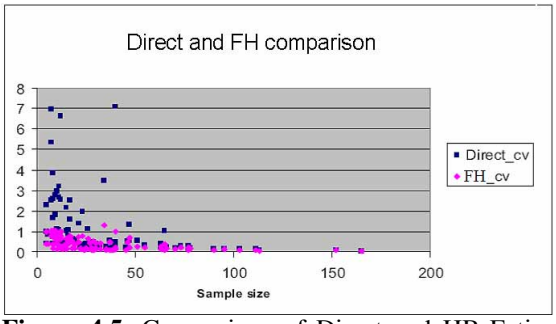

Figure 4.5: Comparison of Direct and HB Estimates (Source: You and Dick 2004)

Figure 4.5 supports the conclusion that the FH approach leads to smoothed estimates, particularly for the domains with relatively small sample sizes. When sample size is small, some direct net undercoverage estimates are negative due to the fact that the overcoverage estimates are larger then the undercoverage estimates. The FH method "corrected" the negative values. All the FH net undercoverage estimates are positive.

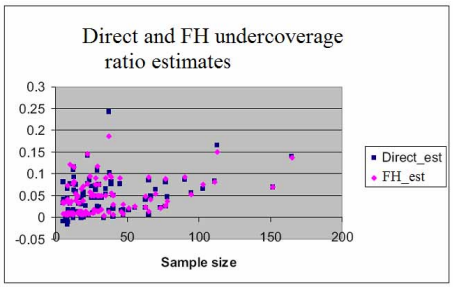

Figure 4.6: Comparison of Direct and FH CVs (Source: You and Dick 2004)

In terms of the CV comparison given in figure 4.6, the HB approach achieves a large CV reduction when the sample sizes are small. As sample size increases, the CV reduction decreases. As the sample size increases, the CVs of the direct and HB estimates quite similar.

4.4 Labour Force Survey

Unemployment rates are produced on a monthly basis in Canada by the Labour Force Survey (LFS). The LFS samples some 53,000 households based on a stratified multi-stage design. The survey reduces response burden by having one-sixth of its sample replaced each month. For a detailed description of the LFS design, see Gambino, Singh, Dufour, Kennedy and Lindeyer (1998). The published provincial and national estimates unemployment rates are a key indicator of economic performance in Canada.

Unemployment rates at levels lower than the provincial level are also of great interest. For instance, the unemployment rates for Census Metropolitan Areas (CMAs, *i.e.*, cities with Population more than 100,000) and Census Agglomerations (CAs, *i.e.*, other urban centers) receive scrutiny at local governments. However, many of the CAs do not have a large enough sample to produce adequate direct estimates. Their estimates need to be produced using SAE techniques. You, Rao and Gambino (2003) used a cross-sectional and time series model to estimate unemployment for such small areas: their methodology borrowed strength both across time and small areas.

Let y_{it} denote the direct LFS estimate of θ_{it} the true unemployment rate of the *ith* CA (small area) at time *t*, for $i = 1, ..., m$, $t = 1, ..., T$, where m is the total number of CAs and T is the (current) time of interest. Assume that the sampling model is

$$y_{it} = \theta_{it} + e_{it}, \quad i = 1,...,m; \quad t = 1,...,T$$

where e_{it} 's are sampling errors. Since the CAs can be treated as strata, the e_{it} 's are uncorrelated between themselves for a given time period *t*. However, the rotation results in a significant level of overlap for the sampled households. This is reflected in the linking model given by $\theta_{it} = x_{it}'\beta + v_i + u_{it}$ where the error structure of the u_{it} 's is assumed to follow an AR(1) process, represented as $u_{it} = u_{i,t-1} + \varepsilon_{it}$; $\varepsilon_{it} \overset{iid}{\sim} (0, \sigma^2)$

The error structure of the e_{it} 's is assumed known, and as this is not the case, the sample based estimates need to be smoothed. You, Rao and Gambino (2003) used the Hierarchical Bayes (HB) procedure to estimate the required parameters in the error and linking equation. They compared numerically three estimators of the unemployment rates in June 1999. These estimators were the direct estimator (Direct Est), a small area estimator based only on the current cross-sectional data (the Fay-Herriot), and one using both the cross-sectional and longitudinal data (Space-time).

Figure 4.7 displays these LFS estimates for the June 1999 unemployment rates for the 62 CAs across Canada. The 62 CAs appear in the order of population size with the smallest CA (Dawson Creek, BC, population is 10,107) on the left and the largest CA (Toronto, Ont., population is 3,746,123) on the right. The Fay-Herriot model tends to shrink the estimates towards the average of the unemployment rates. The space-time model leads to moderate smoothing of the direct LFS estimates. For the CAs with large population sizes and therefore large sample sizes, the direct estimates and the HB estimates are very close to each other; for smaller CAs, the direct and HB estimates differ substantially for some regions.

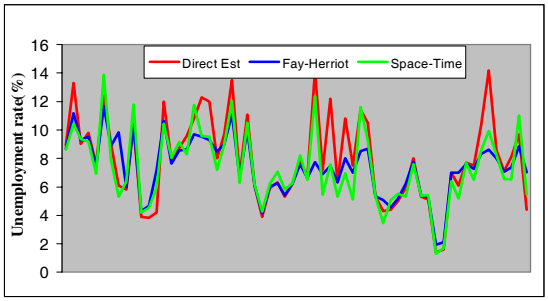

Figure 4.7: Comparison of unemployment rates using Direct, Fay-Herriot, and space-time for June 1999

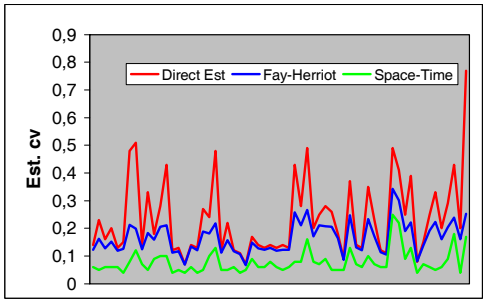

Figure 4.8: Comparison of coefficients of variation of unemployment rates using Direct, Fay-Herriot, and space-time estimates for June 1999

Acknowledgements: The author would like to acknowledge Jon Rao, Peter Dick and Susana Rubin-Bleuer.

References

Australian Bureau of Statistics (2006). A Guide to Small Area Estimation - Version 1.1. Internal ABS document.

Battese, G.E., Harter, R.M., Fuller, W.A. (1988). An Error-Components Model for Prediction of Crop Areas Using Survey and Satellite Data, *Journal of the American Statistical Association,* 83, 28-36.

Brackstone, G. J. (1987). Small area data: policy issues and technical challenges. In R. Platek, J. N. K. Rao, C. E. Sarndal, and M. P. Singh, eds., *Small Area Statistics,* pp. 3-20. John Wiley & Sons, New York.

Béland, Yves Canadian Community Health Survey (2002). Methodological overview. Health report, Statistics Canada, Catalogue no. 82-003-XPE (**0030182-003-XIE.pdf**), Vol. 13, No. 3, ISSN 0840-6529.

Dick, P. (1995). Modelling Net Undercoverage in the 1991 Canadian Census, *Survey Methodology*, 21, 45-54.

Drew, D., Singh, M.P., and Choudhry, G.H. (1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey, *Survey Methodology*, 8, 17-47.

Fay, R.E. and Herriot, R.A. (1979). Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association,* 74, 269-277.

Fuller, W.A. (1999). Environmental Surveys Over Time, *Journal of Agricultural, Biological and Environmental Statistics, 4,* 331-345.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindeyer, J. (1998). *Methodology of the Canadian Labour Force Survey*, Statistics Canada, Catalogue No. 71-526.

Gonzalez, M.E., and Hoza, C. (1978), Small-Area Estimation with Application to Unemployment and Housing Estimates, *Journal of the American Statistical Association,* 73, 7-15.

Hidiroglou M.A. and Singh A., and Hamel M. (2007). some thoughts on small area estimation for the Canadian community health survey (CCHS). Internal Statistics Canada document.

Hidiroglou, M.A. and Särndal, C.E., (1985). Small Domain Estimation: A Conditional Analysis, *Proceedings of the Social Statistics Section, American Statistical Association*, 147-158.

Hidiroglou, M.A. and Patak, Z. (2004). Domain estimation using linear regression. *Survey Methodology*, 30, 67-78.

Levy, P.S. (1971). The Use of Mortality Data in Evaluating Synthetic Estimates*, Proceedings of the Social Statistics Section, American Statistical Association*, pp. 328-331.

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.

Prasad, N.G.N. and Rao, J.N.K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology,* 25, 67-72 .

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

Rao, J.N.K. and Choudhry, H. (1995). Small Area Estimation: Overview and Empirical study. *Business Survey Methods,* Edited by Cox, Binder, Chinnappa, Christianson, Colledge, Kott, Chapter 27.

Rubin-Bleuer, S., Godbout S and Morin Y (2007). Evaluation of small domain estimators for the Canadian Survey of Employment, Payrolls and Hours. Paper presented at the *third International Conference of Establishment Surveys* July 2007 *Statistical of Society Meetings*.

Schaible, W.A. (1978). Choosing Weights for Composite Estimators for Small Area Statistics, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 741-746.

Singh A.C. and Verret F. (2006). Mixed Linear Nonlinear Aggregate level and Matt Type for formulas? Models for Small Area Estimation for Binary count data from Surveys. *Proceedings of the Statistics Canada Symposium.*

Singh, A.C. (2006). Some problems and proposed solutions in developing a small area estimation product for clients. *ASA Proc. Surv. Res. Meth. Sec.*

Singh, M.P., Gambino, J., Mantel, H.J. (1994). Issues and Strategies for Small Area Data, Survey Methodology, 20, 3-22.

Woodruff, R.S. (1966), Use of a Regression Technique to Produce Area Breakdowns of the Monthly National Estimates of Retail Trade, *Journal of the American Statistical Association, 61,* 496-504.

You, Y., and Rao, J.N.K. (2002). A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights, *Canadian Journal of* Statistics, 30, 431-439.

You, Y., Rao, J.N.K. and Dick, J.P. (2002) Benchmarking hierarchical Bayes small area estimators with application in census undercoverage estimation. *Proceedings of the Survey Methods Section 2002, Statistical Society of Canada*, 81 - 86.

You, Y, Rao, J.N.K., and Gambino, J.G. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach, *Survey Methodology,* 29, 25-32.

You, Y and Dick, P. (2004). Hierarchical Bayes Small Area Inference to the 2001 Census Undercoverage Estimation. Proceedings of the ASA Section on Government Statistics, 1836-1840.

.

Appendix: Fay-Herriot computational summary

Description	Computation
1. Model a smooth function of $\overline{Y}_i$	$\theta_i = g\left(\overline{Y}_i\right)$ where $\overline{Y}_i$ is the small area population mean for *i-th* small area; *i=1,…,m*
2. Direct estimate of θ_i	$\hat{\theta}_i = g\left(\hat{\overline{Y}}_i\right)$ where $\hat{\overline{Y}}_i$ is the observed direct estimate
3. Auxiliary data	$\boldsymbol{z}_i = \left(z_{1i}, z_{2i}, \ldots, z_{pi}\right)'$
4. Linking model: Connect the θ_i	$\theta_i = \boldsymbol{z}_i'\boldsymbol{\beta} + v_i$; v_i i.i.d under model $\left(0, \sigma_v^2\right)$; σ_v^2 =model variance
5. Sampling model	$\hat{\theta}_i = \theta_i + e_i$;sampling errors e_i independent $E_p\left(e_i \middle| \theta_i\right) = 0$ and sampling variance $V_p\left(e_i \middle| \theta_i\right) = \psi_i$ (assumed known)
6. Combine 6 and 7	$\hat{\theta}_i = \boldsymbol{z}_i'\boldsymbol{\beta} + v_i + e_i$: Fay-Herriot model
7. Estimation of σ_v^2	**Method of moments**: Solve $h\left(\sigma_v^2\right) = \sum_{i=1}^{m}\left(\hat{\theta}_i - \boldsymbol{z}_i'\tilde{\boldsymbol{\beta}}\left(\sigma_v^2\right)\right)^2 / \left(\psi_i + \sigma_v^2\right) = m - p$ for σ_v^2 via iteration $\sigma_v^{2(r+1)} = \sigma_v^{2(r)} + \left[m - p - h\left(\sigma_v^2\right)\right] \Big/ h_*'\left(\sigma_v^{2(r)}\right)$ constraining to $\sigma_v^{2(r+1)} \geq 0$, where $h_*'\left(\sigma_v^2\right) = -\sum_{i=1}^{m}\left(\hat{\theta}_i - \boldsymbol{z}_i'\tilde{\boldsymbol{\beta}}\right)^2 / \left(\psi_i + \sigma_v^2\right)^2$ is an approximation to the derivative of $h\left(\sigma_v^2\right)$. (see p. 118, Rao (2003))
8. Optimal model-based Fay-Herriot estimator	$\hat{\theta}_i^{FH} = \hat{\gamma}_i\hat{\theta}_i + \left(1-\hat{\gamma}_i\right)\boldsymbol{z}_i'\hat{\boldsymbol{\beta}} = \boldsymbol{z}_i'\hat{\boldsymbol{\beta}} + \hat{\gamma}_i\left(\hat{\theta}_i - \boldsymbol{z}_i'\hat{\boldsymbol{\beta}}\right) = \boldsymbol{z}_i'\hat{\boldsymbol{\beta}} + \hat{v}_i$ where $\hat{\boldsymbol{\beta}}$ is the weighted least squares estimator of $\boldsymbol{\beta}$. Now $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2\right) = \left[\sum_{i=1}^{m}\boldsymbol{z}_i\boldsymbol{z}_i' / \left(\psi_i + \hat{\sigma}_v^2\right)\right]^{-1}\left[\sum_{i=1}^{m}\boldsymbol{z}_i\hat{\theta}_i / \left(\psi_i + \hat{\sigma}_v^2\right)\right]$ where $\hat{\gamma}_i = \hat{\sigma}_v^2 / \left(\psi_i + \hat{\sigma}_v^2\right)$ (see p. 116 Rao (2003))
9. MSE of $\hat{\theta}_{i,FH}$	Leading term of $MSE\left(\hat{\theta}_{i,FH}\right) = E\left(\hat{\theta}_{i,FH} - \theta_i\right)^2$ where the expectation is with respect to the Fay-Herriot model; see step 8; $g_{1i}\left(\sigma_v^2\right) = \gamma_i\psi_i$ shows the efficiency of $\hat{\theta}_{i,FH}$ over direct estimator $\hat{\theta}_i$ is γ_i^{-1} for large number of areas *m*. If $\gamma_i = \sigma_v^2 / \left(\psi_i + \sigma_v^2\right) = 1/2$, then efficiency is 200% or gain in efficiency is 100%.
10. Scenarios for large efficiency gains	Sampling variance ψ_i large or model variance σ_v^2 small relative to ψ_i
11. Nearly unbiased estimator of $MSE\left(\hat{\theta}_{i,FH}\right)$	$mse\left(\hat{\theta}_{i,FH}\right)$: See equation (7.1.26), p. 129, Rao (2003); easily programmable
12. Estimation of small area mean $\overline{Y}_i$	$\hat{\hat{\overline{Y}}}_{i,FH} = g^{-1}\left(\hat{\theta}_{i,FH}\right) = K\left(\hat{\theta}_{i,FH}\right)$
13. MSE estimator of $\hat{\hat{\overline{Y}}}_{i,FH}$	$mse\left(\hat{\hat{\overline{Y}}}_{i,FH}\right) = \left[K'\left(\hat{\theta}_{i,FH}\right)\right]^2 mse\left(\hat{\theta}_{i,FH}\right)$; may not be nearly unbiased. Empirical Bayes (EB) and hierarchical Bayes(HB) methods are better suited for handling non-linear cases , $K\left(\hat{\theta}_{i,FH}\right)$, see p. 133 , Rao (2003)