Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Development of a small area estimation system at Statistics Canada

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

Canada

by Michel A. Hidiroglou, Jean-François Beaumont
and Wesley Yung

Release date: May 7, 2019

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, 2019 (special issue)
Vol. 45, No. 1, pp. 101-126
Statistics Canada, Catalogue No. 12-001-X

Development of a small area estimation system at Statistics Canada

Michel A. Hidiroglou, Jean-François Beaumont and Wesley Yung[1]

Abstract

The demand for small area estimates by users of Statistics Canada's data has been steadily increasing over recent years. In this paper, we provide a summary of procedures that have been incorporated into a SAS based production system for producing official small area estimates at Statistics Canada. This system includes: procedures based on unit or area level models; the incorporation of the sampling design; the ability to smooth the design variance for each small area if an area level model is used; the ability to ensure that the small area estimates add up to reliable higher level estimates; and the development of diagnostic tools to test the adequacy of the model. The production system has been used to produce small area estimates on an experimental basis for several surveys at Statistics Canada that include: the estimation of health characteristics, the estimation of under-coverage in the census, the estimation of manufacturing sales and the estimation of unemployment rates and employment counts for the Labour Force Survey. Some of the diagnostics implemented in the system are illustrated using Labour Force Survey data along with administrative auxiliary data.

Key Words: Small area estimation; Area level model; Unit level model; EBLUP; Hierarchical Bayes methods; Official Statistics.

1 Introduction

Today's data users are becoming more and more sophisticated and are asking for more data and at more detailed levels. For National Statistical Offices (NSOs) facing declining response rates, producing data at finer levels of detail is a particularly daunting challenge. Small area estimation techniques are one way that can be considered to meet this demand to produce estimates for specified sub-populations or small areas. A *small area* refers to a subgroup of the population for which the sample size is so small that direct estimates are not reliable enough to be published. Examples of small areas include a geographical region (e.g., a province, county, municipality, etc.), a demographic group (e.g., age by sex), a demographic group within a geographic region or a detailed industry group. The demand for small area data has been recognized for years (see Brackstone, 1987), but recently, it has greatly increased as noted in the spring 2014 report of the Auditor General of Canada.

The study of small area estimation procedures has a long history at Statistics Canada, beginning in the seventies with Singh and Tessier (1976) and Ghangurde and Singh (1977). Drew, Singh and Choudhry (1982) proposed a sample dependent procedure to estimate employment characteristics below the provincial level. Dick (1995) modeled net undercoverage for the 1991 Canadian Census of Population. The development of a small area estimation system suited to Statistics Canada surveys is well-timed, as there is now a great deal of literature written on the subject, including the books by Rao (2003) and Rao and Molina (2015).

[1] Michel A. Hidiroglou, Business Survey Methods, Statistics Canada, 22th Floor R.H. Coats Building, Ottawa, ON, K1A 0T6, Canada. E-mail: hidirog@yahoo.ca; Jean-François Beaumont, International Cooperation and Corporate Statistical Methods Division, Statistics Canada, 25th Floor R.H. Coats Building, Ottawa, ON, K1A 0T6, Canada; Wesley Yung, Business Survey Methods, Statistics Canada, 22th Floor R.H. Coats Building, Ottawa, ON, K1A 0T6, Canada.

Four papers that have had a great impact in small area estimation (SAE) are Gonzalez and Hoza (1978), Fay and Herriot (1979), Battese, Harter and Fuller (1988), and Prasad and Rao (1990). Gonzalez and Hoza (1978) were among the first to propose small area estimation procedures (mainly synthetic estimation). Fay and Herriot (1979) developed procedures to estimate income for small areas using the long form Census Data. This method and its variants are among the most widely used procedures for producing small area estimates through the integration of auxiliary data with direct survey estimates. Battese et al. (1988) developed a small area procedure to estimate crop areas using survey and satellite data available for individual units. Finally, Prasad and Rao (1990) derived a nearly unbiased estimator of the model-based mean squared error for both the Fay-Herriot and Battese-Harter-Fuller estimators.

The statistical theory of model-based SAE is rather complex and much of the software available at National Statistical Offices has been programmed on a one-time basis and, as such, is not appropriate in a production environment. It was therefore decided to develop a system as it would be beneficial as a production tool, as well as a learning tool for employees. At the time that this was decided, around 2006, there existed computer programs developed by the EURAREA (2004) project for small area estimation. However, this set of programs was no longer in development mode and did not represent the latest advances in small area estimation. Therefore, a flexible small area estimation system that would address the needs of producing small area estimates in production was developed at Statistics Canada. Some of the basic requirements of this small area system included: allowing for both area and unit level models; incorporating the sampling design in the estimation of the parameters of interest and the mean squared error; ensuring that the small area estimates would add up to reliable higher level estimates (i.e., totals), and developing diagnostic tools to test the adequacy of the models used for small area estimation. A prototype system, written in SAS, was therefore developed by Estevao, Hidiroglou and You (2015) to reflect these requirements. This prototype has been transformed into a production system that is currently used by Statistics Canada.

The paper is organized as follows. Section 2 introduces the notation used in the article. Section 3 discusses the options available in the production system for the area level model and Empirical Best Linear Unbiased Prediction (EBLUP) methods. The options for the unit level model with EBLUP methods are presented in Section 4. The Hierarchical Bayes approach is presented in Section 5 for the area level model. Section 6 illustrates the production system using Statistics Canada's Labour Force Survey. Finally, some conclusions are given in Section 7.

2 Core notation and background

We first introduce some notation that will define the various small area estimators included in the production system. Let U denote a population of size N. This population is partitioned into M mutually exclusive and exhaustive areas, where each area $U_i \subset U$, $i = 1, \ldots, M$ has N_i observations. A sample, s, of size n is drawn from the population using a well-defined probability mechanism $p(s)$ and the resulting sample is split into areas $s_i = s \cap U_i$, $i = 1, \ldots, M$. Note that, for some of the areas, the realized sample size n_i may be zero. The set of $m(m \le M)$ areas, where n_i is strictly greater than 0, will be denoted as A. The set of the remaining areas, where n_i is equal than 0, will be denoted as $\bar{A}$.

Let $\pi_j = \sum_{\{s:\ j \in s\}} p(s),\ j \in U,$ be the inclusion probabilities where $\{s:\ j \in s\}$ denotes summation over all samples s containing unit j. We denote the sampling weight for unit j as d_j, where $d_j = \pi_j^{-1}$. The final weight associated with unit j will be denoted as w_j. This weight will normally be the product of the original design weight (d_j) times an adjustment factor that reflects the incorporation of available auxiliary data (via regression or calibration), as well as non-response adjustments. Note that the auxiliary data used in the adjustment factor may not necessarily be the same as those used for small area estimation.

The objective of a small area estimation system is to estimate a population parameter θ_i (e.g., a mean or a total) for each area i for a given variable of interest y when some area sample sizes n_i are too small to use *direct estimation* procedures. A *direct estimator* of θ_i is one that uses values of the variable of interest, y, strictly from the sample units in area i. However, a major disadvantage of such estimators is that unacceptably large standard errors may result: this is especially true if the area sample size is small. Small area procedures use *indirect estimators* that borrow strength across areas, by using models which link all areas through some common parameters. Indirect estimators will be efficient (i.e., increase the effective sample size and thus decrease the standard error) if the model holds for each area. Departures from the model will result in reduced accuracy. There is a wide variety of indirect estimators available and a good summary is provided in Rao and Molina (2015).

Small area estimators are classified as area or unit level depending on the level at which the modeling is performed. *Area level* small area estimators are based on models linking a given parameter of interest to area-specific auxiliary variables. *Unit level* small area estimators are based on models linking the variable of interest to unit-specific auxiliary variables. Area level 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 aggregating them to the appropriate area level. This might be useful in practice because the area level small area estimators may be less prone to outliers than their unit level counterpart.

3 Area level model

The area level small area estimator first appeared in the seminal paper of Fay and Herriot (1979). Following that paper, let the parameter of interest be θ_i; common examples are totals, $Y_i = \sum_{j \in U_i} y_j$, or means, $\bar{Y}_i = Y_i / N_i$. As noted above, the vector of auxiliary variables may differ from the one used in direct estimation and is denoted as $\mathbf{z}$. The area level model can be expressed as two equations.

The first equation, commonly known as the *sampling model*, is given by

$$\hat{\theta}_i = \theta_i + e_i \tag{3.1}$$

and expresses the direct estimate $\hat{\theta}_i$ in terms of the unknown parameter θ_i plus a random error e_i due to sampling. The sampling errors e_i are independently and identically distributed with mean 0 and variance

ψ_i: that is $E_p(e_i \mid \theta_i) = 0$ and $V_p(e_i \mid \theta_i) = \psi_i$, where p denotes expectation in terms of the sample design. Note that ψ_i is also the design variance of $\hat{\theta}_i$ and is typically unknown.

The second equation, known as the *linking model*, is given by

$$\theta_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i \tag{3.2}$$

and expresses the parameter θ_i as a fixed effect $\mathbf{z}_i^T \boldsymbol{\beta}$ plus a random effect v_i multiplied by b_i. In the production system, the b_i term has a default value of one but can be specified by the user to control heteroscedastic errors or the impact of influential observations. The random effects v_i are independently and identically distributed with mean 0 and unknown model variance σ_v^2, that is $E_m(v_i) = 0$ and $V_m(v_i) = \sigma_v^2$ where E_m denotes the model expectation and V_m the model variance. The random errors e_i are independent of the random effects v_i. The combination of the *sampling model* and *linking model* results in a single generalized linear mixed model (GLMM) given by

$$\hat{\theta}_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i + e_i. \tag{3.3}$$

From the Fay-Herriot model (3.3), we observe that $E_{mp}\left(\hat{\theta}_i\right) = \mathbf{z}_i^T \boldsymbol{\beta}$ and $V_{mp}\left(\hat{\theta}_i\right) = b_i^2 \sigma_v^2 + \tilde{\psi}_i$, where $\tilde{\psi}_i = E_m(\psi_i)$ is the smoothed design variance of $\hat{\theta}_i$. In general, we cannot treat ψ_i as fixed, as it is not strictly a function of auxiliary data. If the σ_v^2's and $\tilde{\psi}_i$'s are known, the solution to the GLMM yields the Best Linear Unbiased Predictor (BLUP), $\tilde{\theta}_i^{\text{BLUP}}$

$$\tilde{\theta}_i^{\text{BLUP}} = \begin{cases} \gamma_i \hat{\theta}_i + (1 - \gamma_i)\, \mathbf{z}_i^T \tilde{\boldsymbol{\beta}} & \text{for } i \in A \\ \mathbf{z}_i^T \tilde{\boldsymbol{\beta}} & \text{for } i \in \bar{A} \end{cases} \tag{3.4}$$

where $\gamma_i = (b_i^2 \sigma_v^2) / (\tilde{\psi}_i + b_i^2 \sigma_v^2)$ and $\tilde{\boldsymbol{\beta}} = \left(\sum_{i \in A} \mathbf{z}_i \mathbf{z}_i^T \big/ (\tilde{\psi}_i + b_i^2 \sigma_v^2)\right)^{-1} \sum_{i \in A} \mathbf{z}_i \hat{\theta}_i \big/ (\tilde{\psi}_i + b_i^2 \sigma_v^2)$.

There are four recursive procedures for estimating σ_v^2 and $\boldsymbol{\beta}$ in the production system. The first three assume that $\tilde{\psi}_i$ is known, or that a smoothed version of it is available (see the following section for details). Under this assumption, the variance components can be computed via the Fay-Herriot procedure (FH) as outlined in Fay and Herriot (1979), the restricted maximum likelihood (REML), or the Adjusted Density Maximization (ADM) due to Li and Lahiri (2010). The fourth procedure, WF, due to Wang and Fuller (2003) assumes that ψ_i is estimated by $\hat{\psi}_i$ given that $n_i \geq 2$. The WF procedure does not require any smoothing of the estimated $\hat{\psi}_i$ values before estimating σ_v^2. Wang and Fuller (2003) carried out simulations with n_i ranging from 9 to 36 and found that their procedure yielded reasonable estimates of θ_i and its estimated mean squared error.

The main difference between these four procedures is how the σ_v^2's are computed. They are all based on an iterative scoring algorithm that obtains $\hat{\sigma}_v^2$ as an estimate of the model variance σ_v^2. The FH, REML, and WF procedures may yield $\hat{\sigma}_v^2$'s that are smaller than zero. If this occurs, the $\hat{\sigma}_v^2$'s are set to zero for both the FH and REML procedures. A drawback of truncating the estimated σ_v^2 to zero is that the resulting small area estimator will be synthetic for all areas. Li and Lahiri (2010) suggested the ADM as a way to address the problem of obtaining negative $\hat{\sigma}_v^2$ by maximizing an adjusted likelihood defined as a product of the model variance and a standard likelihood. Although the ADM method always gives a positive solution for σ_v^2, it should be used cautiously because it overestimates the model variance. The REML, FH and ADM procedures use the smoothed values of the estimated $\hat{\psi}_i$ values obtained from the sample or some estimate provided by the user. For the WF procedure, if $\hat{\sigma}_v^2 < 0$, Wang and Fuller (2003) suggested to set $\hat{\sigma}_v^2$ to $0.5\sqrt{\hat{V}(\hat{\sigma}_v^2)}$, where

$$\hat{V}(\hat{\sigma}_v^2) = \sum_{i \in A} 2\kappa_i^2 \left[(\hat{\psi}_i + b_i^2 \hat{\sigma}_v^2)^2 + \frac{(\hat{\psi}_i)^2}{(n_i - 1)} \right]$$

and

$$\kappa_i = \frac{\left[b_i^2 \hat{\sigma}_v^2 + \frac{(n_i + 1)}{(n_i - 1)} \hat{\psi}_i \right]^{-1}}{\sum_{i \in A} \left[b_i^2 \hat{\sigma}_v^2 + \frac{(n_i + 1)}{(n_i - 1)} \hat{\psi}_i \right]^{-1}}.$$

Plugging $\hat{\sigma}_v^2$ and an estimate of $\tilde{\psi}_i$'s into the $\tilde{\theta}_i^{\text{BLUP}}$, defined by equation (3.4), yields the Empirical Best Linear Unbiased Predictor (EBLUP), $\hat{\theta}_i^{\text{EBLUP}}$. It is given by

$$\hat{\theta}_i^{\text{EBLUP}} = \begin{cases} \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) \mathbf{z}_i^T \hat{\boldsymbol{\beta}} & \text{for } i \in A \\ \mathbf{z}_i^T \hat{\boldsymbol{\beta}} & \text{for } i \in \bar{A} \end{cases}$$

where $\hat{\gamma}_i = (b_i^2 \hat{\sigma}_v^2) / (\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^2)$, $\hat{\boldsymbol{\beta}} = \left(\sum_{i \in A} \mathbf{z}_i \mathbf{z}_i^T / (\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^2) \right)^{-1} \sum_{i \in A} \mathbf{z}_i \hat{\theta}_i^{\text{DIR}} / (\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^2)$, and $\ddot{\psi}_i$ is chosen according to the procedure used. For the REML, FH and ADM procedures the $\ddot{\psi}_i$'s are the smoothed values of the estimated $\hat{\psi}_i$ values obtained from the sample or some estimate provided by the user. For the WF procedure, we have that $\ddot{\psi}_i = \hat{\psi}_i$. If the estimated model variance $b_i^2 \hat{\sigma}_v^2$ is relatively small compared with $\ddot{\psi}_i$, then $\hat{\gamma}_i$ will be small and more weight will be attached to the synthetic estimator $\mathbf{z}_i^T \hat{\boldsymbol{\beta}}$. Similarly, more weight is attached to the direct estimator, $\hat{\theta}_i$, if the design variance $\ddot{\psi}_i$ is relatively small.

Details of the required computations can be found in the methodology specifications for the production system in Estevao et al. (2015).

3.1 Estimation of the smooth design variance

The design variance, ψ_i, could be used as an estimator of the smooth design variance $\tilde{\psi}_i = E_m(\psi_i)$ if it were known. In most cases, it is unknown. To get around this difficulty, a design-unbiased variance estimator $\hat{\psi}_i$ of ψ_i is assumed to be available; i.e., $E_p(\hat{\psi}_i) = \psi_i$. Under this assumption, we have that

$$E_{mp}(\hat{\psi}_i) = E_m(\psi_i) = \tilde{\psi}_i.$$

A simple unbiased estimator of the smooth design variance $\tilde{\psi}_i$ is $\hat{\psi}_i$. However, $\hat{\psi}_i$ may be quite unstable when the sample size in domain i is small. A more efficient estimator is obtained by modelling $\hat{\psi}_i$ given $\mathbf{z}_i$. Dick (1995) and Rivest and Belmonte (2000) considered smoothing models given by

$$\log(\hat{\psi}_i) = \mathbf{x}_i^T \boldsymbol{\alpha} + \varepsilon_i,$$

where $\mathbf{x}_i$ is a vector of explanatory variables that are functions of $\mathbf{z}_i$, $\boldsymbol{\alpha}$ is a vector of unknown model parameters to be estimated, and ε_i is a random error with $E_{mp}(\varepsilon_i) = 0$ and constant variance $\sigma_\varepsilon^2 = V_{mp}(\varepsilon_i)$. We also assume that the errors ε_i are identically distributed conditionally on $\mathbf{z}_i$, $i = 1, \ldots, m$. From the above model, we observe that

$$\tilde{\psi}_i = E_{mp}(\hat{\psi}_i) = \exp(\mathbf{x}_i^T \boldsymbol{\alpha})\,\Delta,$$

where $\Delta = E_{mp}(\exp(\varepsilon_i))$. Dick (1995) estimated $\tilde{\psi}_i$ by omitting the factor Δ. Rivest and Belmonte (2000) estimated Δ by assuming that the errors ε_i are normally distributed. However, we observed empirically that the resulting estimator of Δ is sensitive to deviations from the normality assumption. This assumption is avoided by using a method of moments (see Beaumont and Bocci, 2016). This leads to the unbiased estimator of Δ given by

$$\hat{\Delta}(\boldsymbol{\alpha}) = \frac{\sum_{i=1}^m \hat{\psi}_i}{\sum_{i=1}^m \exp(\mathbf{x}_i^T \boldsymbol{\alpha})}.$$

An estimator $\hat{\boldsymbol{\alpha}}$ of the vector of unknown model parameters $\boldsymbol{\alpha}$ is necessary to estimate $\tilde{\psi}_i$. It is obtained using the ordinary least squares method as

$$\hat{\boldsymbol{\alpha}} = \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_{i=1}^m \mathbf{x}_i \log(\hat{\psi}_i).$$

The estimator $\hat{\hat{\psi}}_i$ of $\tilde{\psi}_i$ is then given by

$$\hat{\hat{\psi}}_i = \exp(\mathbf{x}_i^T \hat{\boldsymbol{\alpha}})\,\hat{\Delta}(\hat{\boldsymbol{\alpha}}).$$

A nice property of $\hat{\hat{\psi}}_i$ is that the average of the smooth design variance estimator, $\hat{\hat{\psi}}_i$, is equal to the average of the direct variance estimator, $\hat{\psi}_i$; i.e.,

$$\frac{\sum_{i=1}^m \hat{\hat{\psi}}_i}{m} = \frac{\sum_{i=1}^m \hat{\psi}_i}{m}.$$

This ensures that $\hat{\hat{\psi}}_i$ does not systematically overestimate or underestimate $\tilde{\psi}_i = E_{mp}(\hat{\psi}_i)$.

3.2 Benchmarking

If the parameter of interest θ_i is a total $(\theta_i = Y_i)$, the user may wish to have the sum of the small area estimates, $\hat{\theta} = \sum_{i \in A \cup \bar{A}} \hat{\theta}_i^{\text{EBLUP}}$, agree with the estimated totals $\hat{Y} = \sum_{i \in A} \hat{Y}_i$ at the overall sample level s;

i.e., $\hat{\theta} = \hat{Y}$. In the case of a mean, $\theta_i = \bar{Y}_i$, this benchmarking condition becomes $\sum_{i \in A \cup \bar{A}} N_i \hat{\theta}_i^{\text{EBLUP}} = \sum_{i \in A} N_i \hat{\theta}_i$, where $\hat{\theta}_i = \hat{\bar{Y}}_i$.

Two methods are available in the production system to ensure benchmarking for area based small area estimates. The first one is based on a difference adjustment and the second one is based on an augmented vector. They are valid for any method used to compute $\hat{\theta}_i^{\text{EBLUP}}$ or whether the variance estimate $\ddot{\psi}_i$ has been smoothed or not. The benchmarking based on a difference adjustment is an adaptation of the benchmarking given in Battese et al. (1988). The benchmarking based on an augmented vector is due to Wang, Fuller and Qu (2008).

Difference adjustment: For this method, the $\hat{\theta}_i^{\text{EBLUP}}$ estimator is adjusted only for those areas where the realized sample size $n_i \geq 1$, $i \in A$ and the synthetic estimates $\mathbf{z}_i^T \hat{\boldsymbol{\beta}}$ for $i \in \bar{A}$ are left as is. The resulting benchmarked estimator is given by $\hat{\theta}_i^{\text{EBLUP},\, b}$ and is defined as follows

$$\hat{\theta}_i^{\text{EBLUP},\, b} = \begin{cases} \hat{\theta}_i^{\text{EBLUP}} + \alpha_i \left(\hat{\theta}^* - \sum_{d \in A} \omega_d \hat{\theta}_d^{\text{EBLUP}} \right) & \text{for } i \in A \\ \mathbf{z}_i^T \hat{\boldsymbol{\beta}} & \text{for } i \in \bar{A} \end{cases}$$

where $\alpha_i = \left\{ \sum_{i \in U_A} \omega_i^2 \left(\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^2 \right) \right\}^{-1} \omega_i \left(\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^2 \right)$ for $i \in A$, $\omega_i = 1$, if the benchmarking is to a total, and $\omega_i = N_i / N$, if the benchmarking is for the mean. The estimator $\hat{\theta}^*$ is a value provided by the user that represents the total or mean of the y-values of population U. The benchmarking ensures that $\sum_{i \in A \cup \bar{A}} \omega_i \hat{\theta}_i^{\text{EBLUP},\, b} = \hat{\theta}^*$.

Augmented vector: The vector $\mathbf{z}_i^T$ is augmented with $\omega_i \ddot{\psi}_i$, to form $\mathbf{z}_i^{*T} = \left(\mathbf{z}_i^T, \omega_i \ddot{\psi}_i \right)$ with ω_i and $\ddot{\psi}_i$ as previously defined. The resulting augmented generalized linear mixed model (GLMM) equation is given by

$$\hat{\theta}_i = \mathbf{z}_i^{*T} \boldsymbol{\beta}^* + b_i v_i^* + e_i \tag{3.5}$$

where $E_m \left(v_i^* \right) = 0$ and $V_m \left(v_i^* \right) = \sigma_v^{*2}$. The estimates for $\boldsymbol{\beta}^*$ and σ_v^{*2} are once more solved recursively for the four EBLUP procedures that we denote as $\hat{\theta}_i^{\text{EBLUP*}}$.

The resulting benchmarked estimator $\hat{\theta}_i^{\text{EBLUP*},\, b}$ is given by

$$\hat{\theta}_i^{\text{EBLUP*},\, b} = \begin{cases} \hat{\gamma}_i^* \hat{\theta}_i^{\text{EBLUP*}} + \left(1 - \hat{\gamma}_i^* \right) \mathbf{z}_i^{*T} \hat{\boldsymbol{\beta}}^* & \text{for } i \in A \\ \mathbf{z}_i^{*T} \, \hat{\boldsymbol{\beta}}^* & \text{for } i \in \bar{A} \end{cases}$$

where $\hat{\gamma}_i^* = \left(b_i^2 \hat{\sigma}_v^{*2} \right) / \left(\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^{*2} \right)$, and $\hat{\boldsymbol{\beta}}^* = \left(\sum_{i \in A} \mathbf{z}_i^* \mathbf{z}_i^{*T} \Big/ \left(\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^{*2} \right) \right)^{-1} \sum_{i \in A} \mathbf{z}_i^* \hat{\theta}_i \Big/ \left(\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^{*2} \right)$.

All the components of $\hat{\theta}_i^{\text{EBLUP*},\, b}$ are computed using the augmented model given by (3.5). It can be shown that $\sum_{i \in A \cup \bar{A}} \omega_i \hat{\theta}_i^{\text{EBLUP*},\, b} = \sum_{i \in A} \omega_i \hat{\theta}_i$, and hence the benchmarking holds.

The difference adjustment and augmented vector methods are two ways that benchmarking can be satisfied. Wang et al. (2008) suggested other procedures that can be used. Specifically, they adapted the self-calibrated estimator You and Rao (2002) developed in the context of the unit level model to the area level model. You, Rao and Hidiroglou (2013) obtained an estimator of the mean squared prediction error and its bias under a misspecified model.

3.3 Mean squared error estimation

The reliability of the EBLUP estimators is obtained as $\text{MSE}\left(\hat{\theta}_i^{\text{EBLUP}}\right) = E\left(\hat{\theta}_i^{\text{EBLUP}} - \theta_i\right)^2$. The expectation is with respect to models (3.3) for the non-benchmarked estimator, and (3.5) for the benchmarked estimator.

The estimated Mean Squared Errors (MSEs) of the area level estimators are given in Table 3.1. The specific form of the g terms and the estimated variances can be found in Rao and Molina (2015) or in Estevao et al. (2015). For the benchmarked estimators, the estimated MSE for the difference adjustment approach uses the non-benchmarked MSE formulas. For the case of the augmented vector approach, the MSE is based on augmenting the vector $\mathbf{z}_i^T$ with $\omega_i \ddot{\psi}_i$.

Table 3.1
MSE estimates (mse) for the area level estimators

Estimator	mse
Fay-Herriot	$\text{mse}\left(\hat{\theta}_i^{\text{FH}}\right) = \begin{cases} g_{0i} + g_{1i} + g_{2i} + 2g_{3i} & \text{for } i \in A \\ \mathbf{z}_i^T \text{ var}\left(\hat{\boldsymbol{\beta}}\right) \mathbf{z}_i + b_i^2 \hat{\sigma}_v^2 & \text{for } i \in \bar{A} \end{cases}$
ADM	$\text{mse}\left(\hat{\theta}_i^{\text{ADM}}\right) = \begin{cases} g_{0i} + g_{1i} + g_{2i} + 2g_{3i} & \text{for } i \in A \\ \mathbf{z}_i^T \text{ var}\left(\hat{\boldsymbol{\beta}}\right) \mathbf{z}_i + b_i^2 \hat{\sigma}_v^2 & \text{for } i \in \bar{A} \end{cases}$
REML	$\text{mse}\left(\hat{\theta}_i^{\text{REML}}\right) = \begin{cases} g_{1i} + g_{2i} + 2g_{3i} & \text{for } i \in A \\ \mathbf{z}_i^T \text{var}\left(\hat{\boldsymbol{\beta}}\right) \mathbf{z}_i + b_i^2 \hat{\sigma}_v^2 & \text{for } i \in \bar{A} \end{cases}$
WF	$\text{mse}\left(\hat{\theta}_i^{\text{WF}}\right) = \begin{cases} g_{1i} + g_{2i} + 2g_{3i} + g_{4i} & \text{for } i \in A \\ \mathbf{z}_i^T \text{var}\left(\hat{\boldsymbol{\beta}}\right) \mathbf{z}_i + b_i^2 \hat{\sigma}_v^2 & \text{for } i \in \bar{A} \end{cases}$

The various g terms in Table 3.1 can be interpreted as follows. The g_{0i} is a bias correction term for FH and ADM. The g_{1i} term given by $g_{1i} = \hat{\gamma}_i \ddot{\psi}_i$, accounts for most of the MSE if the number of areas is large. The g_{2i} term accounts for the estimation of $\boldsymbol{\beta}$, and $2g_{3i}$ accounts for the estimation of σ_v^2. The g_{4i} term in the WF procedure reflects that the estimated value of ψ_i, $\hat{\psi}_i$, has been used. The estimated variance of $\hat{\boldsymbol{\beta}}$, given by $\text{var}\left(\hat{\boldsymbol{\beta}}\right) = \left(\sum_{i \in A} \frac{\mathbf{z}_i \mathbf{z}_i^T}{\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^2}\right)^{-1}$ is dependent on the particular procedure used to estimate σ_v^2.

4 Unit level model

The original unit level model was proposed by Battese et al. (1988). They assumed the following nested error model

$$y_{ij} = \mathbf{z}_{ij}^{T}\boldsymbol{\beta} + v_i + e_{ij} \quad \text{for} \quad i = 1, \ldots, m \quad \text{and} \quad j \in U_i \tag{4.1}$$

where $v_i \overset{\text{ind}}{\sim} (0, \sigma_v^2)$ are the random effects and are independent of the random errors, e_{ij}, with $e_{ij} \overset{\text{ind}}{\sim} (0, \sigma_e^2)$. The production system includes a slight modification to the error structure of the random errors. That is, $e_{ij} \overset{\text{ind}}{\sim} \left(0, \sigma_e^2 \big/ a_{ij}\right)$, where $a_{ij} > 0$ are positive constants that account for heteroscedasticity.

The production system computes small area estimates for means $\left(\bar{Y}_{ic} = \sum_{j \in U_i} c_{ij} y_{ij} \big/ \sum_{j \in U_i} c_{ij}\right)$ and totals $\left(Y_{ic} = \sum_{j \in U_i} c_{ij} \bar{Y}_i\right)$. The c_{ij} values are fixed positive constants known for all population units. The addition of c_{ij} was necessary to allow the use of the system by some business surveys conducted at Statistics Canada (see Rubin-Bleuer, Jang and Godbout, 2016). The available auxiliary data are either totals $\mathbf{Z}_{ic} = \sum_{j \in U_i} c_{ij} \mathbf{z}_{ij}$, or means $\bar{\mathbf{Z}}_{ic} = \sum_{j \in U_i} c_{ij} \mathbf{z}_{ij} \big/ \sum_{j \in U_i} c_{ij}$.

In what follows, we provide the estimators of the population means $\bar{Y}_{ic}$, say $\hat{\theta}_i^{\text{SAE}}$, where $i = 1, \ldots, M$. Estimates of the corresponding totals Y_{ic}, are obtained by multiplying $\hat{\theta}_i^{\text{SAE}}$ by $\sum_{j=1}^{N_i} c_{ij}$.

The design weighted sample mean of the y's and $\mathbf{z}$'s are respectively

$$\bar{y}_{iwc} = \left(\sum_{j \in s_i} w_{ij} c_{ij}\right)^{-1} \sum_{j \in s_i} w_{ij} c_{ij} y_{ij}$$

and

$$\bar{\mathbf{z}}_{iwc} = \left(\sum_{j \in s_i} w_{ij} c_{ij}\right)^{-1} \sum_{j \in s_i} w_{ij} c_{ij} \mathbf{z}_{ij}.$$

The model based weighted means are

$$\bar{y}_{ia} = \left(\sum_{j \in s_i} a_{ij}\right)^{-1} \left(\sum_{j \in s_i} a_{ij} y_{ij}\right)$$

and

$$\bar{\mathbf{z}}_{ia} = \left(\sum_{j \in s_i} a_{ij}\right)^{-1} \left(\sum_{j \in s_i} a_{ij} \mathbf{z}_{ij}\right).$$

Battese et al. (1988) did not include survey design weights in their procedure, thereby forsaking design consistency unless the design was self-weighting. We refer to this estimator as EBLUP $\left(\hat{\theta}_i^{\text{EBLUP}}\right)$. However, EBLUP is the most efficient estimator under model (4.1), with error structure $e_{ij} \overset{\text{ind}}{\sim} \left(0, \sigma_e^2 \big/ a_{ij}\right)$, and this is the reason that it is included in the production system.

Kott (1989), Prasad and Rao (1999), and You and Rao (2002) proposed the use of design-consistent model based estimators for the area means by including the survey weight. The You and Rao (2002) procedure was suitably modified to reflect the heteroscedastic residuals and the c_{ij}'s. The resulting Pseudo-EBLUP estimator, denoted as PEBLUP $\left(\hat{\theta}_i^{\text{PEBLUP}}\right)$, was included in the production system as it is design consistent.

The EBLUP estimator is defined as

$$\hat{\theta}_i^{\text{EBLUP}} = \begin{cases} \hat{\gamma}_{ia} \bar{y}_{ia} + \left(\bar{\mathbf{Z}}_{ic} - \hat{\gamma}_{ia} \bar{\mathbf{z}}_{ia}\right)^{T} \hat{\boldsymbol{\beta}}^{\text{EBLUP}} & \text{if } i \in A \\ \bar{\mathbf{Z}}_{ic}^{T} \hat{\boldsymbol{\beta}}^{\text{EBLUP}} & \text{if } i \in \bar{A} \end{cases}$$

where $\hat{\gamma}_{ia} = \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \middle/ \sum_{j \in s_i} a_{ij}\right)^{-1} \hat{\sigma}_v^2$. The terms $\bar{y}_{ia}$ and $\bar{\mathbf{z}}_{ia}$, are the previously defined model based weighted means for y and $\mathbf{z}$ respectively. The regression vector $\boldsymbol{\beta}$ is estimated as

$$\hat{\boldsymbol{\beta}}^{\text{EBLUP}} = \left(\sum_{i=1}^{m} \sum_{j \in s_i} a_{ij} c_{ij} \left(\mathbf{z}_{ij} - \hat{\gamma}_{iac} \bar{\mathbf{z}}_{iac}\right) \mathbf{z}_{ij}^T\right)^{-1} \sum_{i=1}^{m} \sum_{j \in s_i} a_{ij} c_{ij} \left(\mathbf{z}_{ij} - \hat{\gamma}_{iac} \bar{\mathbf{z}}_{iac}\right) y_{ij}.$$

The PEBLUP estimator, $\hat{\theta}_i^{\text{PEBLUP}}$, is given by

$$\hat{\theta}_i^{\text{PEBLUP}} = \begin{cases} \hat{\gamma}_{iwc} \bar{y}_{iwc} + \left(\bar{\mathbf{Z}}_{ic} - \hat{\gamma}_{iwc}\, \bar{\mathbf{z}}_{iwc}\right)^T \hat{\boldsymbol{\beta}}^{\text{PEBLUP}} & \text{if } i \in A \\ \bar{\mathbf{Z}}_{ic}^T \hat{\boldsymbol{\beta}}^{\text{PEBLUP}} & \text{if } i \in \bar{A} \end{cases}$$

where $\hat{\gamma}_{iwc} = (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \delta_{iwc}^2)^{-1} (\hat{\sigma}_v^2)$, and $\delta_{iwc}^2 = \left(\sum_{j \in s_i} w_{ij} c_{ij}\right)^{-2} \left(\sum_{j \in s_i} \left(w_{ij} c_{ij}\right)^2 \middle/ a_{ij}\right)$. The terms $\bar{y}_{iwc}$ and $\bar{\mathbf{z}}_{iwc}$, are the previously defined design based weighted means for y and $\mathbf{z}$ respectively. The regression vector $\boldsymbol{\beta}$ is estimated as

$$\hat{\boldsymbol{\beta}}^{\text{PEBLUP}} = \left(\sum_{i=1}^{m} \sum_{j \in s_i} w_{ij} a_{ij} \left(\mathbf{z}_{ij} - \hat{\gamma}_{iwa} \bar{\mathbf{z}}_{iwa}\right) \mathbf{z}_{ij}^T\right)^{-1} \sum_{i=1}^{m} \sum_{j \in s_i} w_{ij} a_{ij} \left(\mathbf{z}_{ij} - \hat{\gamma}_{iwa} \bar{\mathbf{z}}_{iwa}\right) y_{ij}$$

where $\bar{\mathbf{z}}_{iwa} = \left(\sum_{j \in s_i} w_{ij} a_{ij}\right)^{-1} \sum_{j \in s_i} w_{ij} a_{ij} \mathbf{z}_{ij}$, $\hat{\gamma}_{iwa} = (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \delta_{iwa}^2)^{-1} \hat{\sigma}_v^2$ and with δ_{iwa}^2 computed as $\delta_{iwa}^2 = \left(\sum_{j \in s_i} w_{ij} a_{ij}\right)^{-2} \left(\sum_{j \in s_i} (w_{ij} a_{ij})^2 \middle/ a_{ij}\right)$.

The components of variance, σ_e^2 and σ_v^2, are estimated using the fitting-of-constants (not weighted by the survey weights) method, as given by Battese et al. (1988) or Rao (2003). The resulting estimators of σ_e^2 are always greater than or equal to zero, but the estimator of σ_v^2 may be negative. If $\hat{\sigma}_v^2 < 0$, it is set to zero, implying that there are no area effects. The associated estimated MSEs are obtained by extending You and Rao (2002) and Stukel and Rao (1997).

Note that if the sample s is selected from the universe U, the realized sampling fraction, $f_i = n_i / N_i$, could be non-negligible. For estimating a population mean, $\bar{Y}_i$, Rao and Molina (2015), accounted for non-negligible sampling fractions by expressing it as

$$\bar{Y}_i = f_i \bar{y}_{is} + (1 - f_i)\, \bar{y}_{i\bar{s}}$$

where $\bar{y}_{is}$ is the sample mean of the i^{th} sampled area and $\bar{y}_{i\bar{s}}$ is the sample mean of the non-sampled units within that area. They predicted $\bar{y}_{i\bar{s}}$ using the unit level model given by equation (4.1). Their expressions correspond to the case when $c_{ij} = 1$. This estimator was extended by Rubin-Bleuer (2014) to include the EBLUP and PEBLUP estimators for the case that c_{ij} is arbitrary. Specific details that also account for MSE estimation can be found in Estevao et al. (2015).

4.1 Benchmarking

The current production system does not have a procedure to benchmark the estimates obtained via the unit level model. However, the difference adjustment approach can be suitably modified to allow this. The EBLUP and PEBLUP estimators are of the form

$$\hat{\theta}_i^{\text{SAE}} = \begin{cases} \hat{\gamma}_i^* \bar{y}_i^* + (\bar{\mathbf{Z}}_{ic} - \hat{\gamma}_i^* \bar{\mathbf{z}}_i^*)^T \hat{\boldsymbol{\beta}}^{\text{SAE}} & \text{if } i \in A \\ \bar{\mathbf{Z}}_{ic}^T \hat{\boldsymbol{\beta}}^{\text{SAE}} & \text{if } i \in \bar{A} \end{cases}$$

where $\hat{\gamma}_i^*$, $\bar{y}_i^*$, $\bar{\mathbf{z}}_i^*$, and $\hat{\boldsymbol{\beta}}^{\text{SAE}}$ correspond to the terms defined previously: $\hat{\gamma}_i^*$ is equal to $\hat{\gamma}_{ia}$ for EBLUP, and to $\hat{\gamma}_{iwc}$ for PEBLUP; $\bar{y}_i^*$ is equal to $\bar{y}_{ia}$ for EBLUP, and to $\bar{y}_{iwc}$ for PEBLUP; $\bar{\mathbf{z}}_i^*$ is equal to $\bar{\mathbf{z}}_{ia}$ for EBLUP, and to $\bar{\mathbf{z}}_{iwc}$ for PEBLUP; and, $\hat{\boldsymbol{\beta}}^{\text{SAE}}$ is equal to $\hat{\boldsymbol{\beta}}^{\text{EBLUP}}$ for EBLUP, and to $\hat{\boldsymbol{\beta}}^{\text{PEBLUP}}$ for PEBLUP.

Suppose that $\hat{\theta}_i^{\text{SAE}}$ needs to be benchmarked to $\hat{\theta}^*$. The corresponding benchmarked estimator is

$$\hat{\theta}_i^{\text{SAE},\,b} = \begin{cases} \hat{\theta}_i^{\text{SAE}} + \alpha_i \left(\theta^* - \sum_{d \in A} \omega_d \hat{\theta}_d^{\text{SAE}} \right) & \text{if } \; i \in A \\ \bar{\mathbf{Z}}_{ic}^T \hat{\boldsymbol{\beta}}^{\text{SAE}} & \text{if } \; i \in \bar{A} \end{cases}$$

where $\alpha_i = \left(\sum_{d \in A} \omega_d^2 \tau_d \right)^{-1} (\omega_i \tau_i)$. The ω_i term is defined as follows: $\omega_i = 1$ if the benchmarking is to a total and $\omega_i = N_i / N$ if the benchmarking is for the mean. Possible choices of the τ_i's are $\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \delta_{ia}^2$, $\delta_{ia}^2 = \left(\sum_{j=1}^{n_i} a_{ij} \right)^{-1}$, for EBLUP, and $\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \delta_{iwc}^2$ for PEBLUP.

4.2 Mean squared error estimation

The mean squared error estimates of the unit level estimators are based on estimating its mean squared error, given model (4.1) and error structure $e_{ij} \overset{\text{ind}}{\sim} \left(0,\, \sigma_e^2 / a_{ij}\right)$. Table 4.1 displays these estimated MSE's.

Table 4.1
MSE estimates for the unit level estimators

Estimator	mse
EBLUP	$\text{mse}\left(\hat{\theta}_i^{\text{EBLUP}}\right) = \begin{cases} g_{1ia} + g_{2ia} + 2g_{3ia} & \text{for } \; i \in A \\ \bar{\mathbf{Z}}_i^T \text{ var}\left(\hat{\boldsymbol{\beta}}^{\text{EBLUP}}\right) \bar{\mathbf{Z}}_i + \hat{\sigma}_v^2 & \text{for } \; i \in \bar{A} \end{cases}$
PEBLUP	$\text{mse}\left(\hat{\theta}_i^{\text{PEBLUP}}\right) = \begin{cases} g_{1iw} + g_{2iw} + 2g_{3iw} & \text{for } \; i \in A \\ \bar{\mathbf{Z}}_i^T \text{ var}\left(\hat{\boldsymbol{\beta}}^{\text{PEBLUP}}\right) \bar{\mathbf{Z}}_i + \hat{\sigma}_v^2 & \text{for } \; i \in \bar{A} \end{cases}$

The various g terms in Table 4.1 can be interpreted in a similar way to those associated with the area level MSE's. The g_{1i}'s are denoted as g_{1ia} for EBLUP, and g_{1iw} for PEBLUP account for most of the MSE if the number of areas is large. The g_{2i}'s account for the estimation of $\boldsymbol{\beta}$, and the $2g_{3i}$'s account for the estimation of σ_v^2 and σ_e^2.

The estimated variances of $\hat{\boldsymbol{\beta}}^{\text{EBLUP}}$ and $\hat{\boldsymbol{\beta}}^{\text{PEBLUP}}$ are respectively given by

$$\text{var}\left(\hat{\boldsymbol{\beta}}^{\text{EBLUP}}\right) = \hat{\sigma}_e^2 \left(\sum_{i \in A} \sum_{j \in s_i} a_{ij} \left(\mathbf{z}_{ij} - \hat{\gamma}_{ia} \bar{\mathbf{x}}_{ia} \right) \mathbf{z}_{ij}^T \right)^{-1}$$

and

$$\operatorname{var}\left(\hat{\boldsymbol{\beta}}^{\text{PEBLUP}}\right) = \hat{\sigma}_e^2 \left(\sum_{i\in A}\sum_{j\in s_i} \mathbf{z}_{ij}^* \mathbf{z}_{ij}^{*T}\right)^{-1} \left(\sum_{i\in A}\sum_{j\in s_i} \mathbf{z}_{ij}^* \mathbf{z}_{ij}^{*T} \Big/ a_{ij}\right) \left(\sum_{i\in A}\sum_{j\in s_i} \mathbf{z}_{ij}^* \mathbf{z}_{ij}^{*T}\right)^{-1}$$

where $\mathbf{z}_{ij}^* = w_{ij} a_{ij} \left(\mathbf{z}_{ij} - \hat{\gamma}_{iwa} \overline{\mathbf{z}}_{iwa}\right)$.

The specific form of the g terms and the estimated variances can be found in Estevao et al. (2015).

5 Hierarchical Bayes (HB) method

The basic Fay-Herriot area level model includes a linear sampling model for direct survey estimates and a linear linking model for the parameters of interest. Such models are *matched* because θ_i appears as a linear function in both the sampling and linking models. There are instances when these equations are not matched such as when a function, $h(\theta_i)$, is modelled as a linear function of explanatory variables instead of θ_i. The *sampling model* and *linking model* pair is

$$\hat{\theta}_i = \theta_i + e_i \tag{5.1}$$

and

$$h(\theta_i) = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i \tag{5.2}$$

where $e_i \overset{\text{ind}}{\sim} N(0, \psi_i)$ and $v_i \overset{\text{ind}}{\sim} N(0, \sigma_v^2)$.

The model pair given by (5.1) and (5.2) is referred to as an *unmatched* model. Nonlinear linking models are often needed in practice to provide a better model fit to the data. For example, if the parameter of interest is a probability or a rate within the range of 0 and 1, a linear linking model with normal random effects may not be appropriate. A linking model, in this case, could be a logistic or log-linear model. Such a model was used to adjust counts for detailed levels for the 2011 Census of Canada. A good description of what is involved to carry out such an adjustment can be found in Dick (1995) and You, Rao and Dick (2004).

The production system includes the following choices of $h(\theta_i)$

$$h(\theta_i) = \begin{cases} \theta_i & : \text{ Matched Fay-Herriot (FH) model} \\ \log(\theta_i) & : \text{ Unmatched log-linear model} \\ \log(\theta_i/(\theta_i + C_i)) & : \text{ Unmatched log census undercount model.} \end{cases} \tag{5.3}$$

The inclusion of $h(\theta_i) = \theta_i$ corresponds to the matched model represented by equations (3.1) and (3.2). An advantage of choosing the Hierarchical Bayes method is that the estimated σ_v^2 cannot be negative. The function $\log(\theta_i)$, where θ_i is equal to the population mean $\overline{Y}_i$, was used in Fay and Herriot (1979). Their context was to estimate per capita income (PCI) for small places in the United States with a population less than 1,000. The function $h(\theta_i)$, $\log(\theta_i/(\theta_i + C_i))$, was included to support the methodology to estimate the net undercoverage in Canadian Censuses. In this model, θ_i represents the number of individuals not counted in the census, while C_i is the known census count. As a result, $\theta_i/(\theta_i + C_i)$ is the proportion of individuals undercounted by the Census.

The sampling variances, ψ_i, are assumed known for all the linking models represented by (5.2). The variances are assumed to be estimated for the first two functions (the matched Fay-Herriot and unmatched log-linear model) given in (5.3). If the sampling variances, ψ_i, are assumed known, then the unknown parameters in the sampling model (5.1) and the linking model (5.2) can be presented in a hierarchical Bayes (HB) framework as follows

$$\left[\hat{\theta}_i \,|\, \theta_i\right] \sim N\left(\theta_i,\, \psi_i\right), \quad i = 1, \ldots, m$$

and

$$[h(\theta_i) \,|\, \beta,\, \sigma_v^2] \sim N(\mathbf{z}_i^T \boldsymbol{\beta},\, b_i^2 \sigma_v^2).$$

If the sampling variances are unknown, they are estimated by adding

$$[d_i \widehat{\psi}_i \,|\, \psi_i] \sim \psi_i \chi_{d_i}^2$$

where $\chi_{d_i}^2$ follows a chi-square distribution with $d_i = (n_i - 1)$ degrees of freedom.

The model parameters $\boldsymbol{\beta}$, σ_v^2 and ψ_i (when it is unknown) are assumed to obey prior distributions. The distributions used in the production system for $\boldsymbol{\beta}$ and σ_v^2 are the flat prior, $\pi(\boldsymbol{\beta}) \propto 1$, and $\pi(\sigma_v^2) \propto (\sigma_v^2)^{-1/2}$. If ψ_i is estimated, the prior $\pi(\psi_i) \propto (\psi_i)^{-1/2}$ is added to the Bayesian model. These prior distributions are multiplied by the density functions of the distributions associated with the sampling and linking models. This yields a joint likelihood function in terms of the model parameters. This function is used to obtain a full conditional (posterior) distribution for each of the unknown parameters. For some of these, the resulting distribution has a tractable or well-known form. For others, the resulting distribution is a product of density functions with no known form. All HB methods involve estimation of the model parameters through repeated sampling of their respective full conditional distributions.

Markov Chain Monte Carlo (MCMC) methods are used to obtain estimates from the full conditional distribution of each parameter. Gibbs sampling is used repeatedly to sample from the full conditional distributions. The Gibbs sampling method (Gelfand and Smith, 1990) with the Metropolis-Hastings algorithm (Chib and Greenberg, 1995) are used to find the posterior means and posterior variances; see Estevao et al. (2015) for details. The various estimators of θ_i resulting from (5.3) are denoted as $\hat{\theta}_i^{\mathrm{HB}}$.

5.1 Benchmarked HB estimator

Benchmarking of the estimators uses the *difference adjustment method* described in Section 3.2. That is, the benchmarked estimators $\hat{\theta}_i^{\mathrm{HB}}$ are computed as

$$\hat{\theta}_i^{\mathrm{HB},\, b} = \begin{cases} \hat{\theta}_i^{\mathrm{HB}} + \alpha_i \left(\hat{\theta}^* - \sum_{d \in A} \omega_d \hat{\theta}_d^{\mathrm{HB}}\right) & \text{for } \; i \in A \\ \mathbf{z}_i^T \hat{\boldsymbol{\beta}} & \text{for } \; i \in \overline{A} \end{cases}$$

where $\alpha_i = \left(\sum_{i \in A} \omega_i^2 \left(\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^{2\mathrm{HB}}\right)\right)^{-1} \omega_i \left(\ddot{\psi}_i + b_i^2 \hat{\sigma}_v^{2\mathrm{HB}}\right)$ for $i \in A$, and $\hat{\theta}^*$ is the benchmark value. The terms ω_i are defined as follows: $\omega_i = 1$ if the benchmarking is to a total, and $\omega_i = N_i / N$ if the benchmarking is for the mean. The $\ddot{\psi}_i$'s are either known or unknown. The $\hat{\theta}^*$ can be a value provided by the user that represents the total or mean of the y-values of population U. The benchmarking ensures that $\sum_{i \in A \cup \bar{A}} \omega_i \hat{\theta}_i^{\text{HB},\, b} = \hat{\theta}^*$.

6 Application to Labour Force Survey (LFS) data

Statistics Canada's LFS is a monthly survey with a stratified two-stage design. It is designed to produce reliable unemployment rate estimates for the 55 Employment Insurance Economic Regions (EIER) in Canada. The unemployment rate in any given area i is defined as the ratio

$$\theta_i = \frac{\sum_{j \in U_i} y_{1j}}{\sum_{j \in U_i} y_{2j}},$$

where y_{1j} is a binary variable indicating whether person j is unemployed $(y_{1j} = 1)$ or not $(y_{1j} = 0)$, and y_{2j} is a binary variable indicating whether person j is in the labour force $(y_{2j} = 1)$ or not $(y_{2j} = 0)$. The direct estimator of θ_i is the calibration composite estimator described in Fuller and Rao (2001). See also Singh, Kennedy and Wu (2001) and Gambino, Kennedy and Singh (2001). It can be written in the weighted form

$$\hat{\theta}_i = \frac{\sum_{j \in s_i} w_j y_{1j}}{\sum_{j \in s_i} w_j y_{2j}},$$

where w_j is a calibration composite weight for person j.

As mentioned above, the calibration composite estimator is reliable for the estimation of the unemployment rate for the 55 EIERs. There is also interest in obtaining reliable estimates for 149 areas (cities) in Canada. Among them, there are 34 Census Metropolitan Areas (CMA) and 115 Census Areas (CA). The CMAs are the largest cities in terms of population size and they usually have a large sample size as well. Some of the CAs have a very small sample size, sometimes even 0. For those CAs and other larger CAs, the sample size is not large enough to produce sufficiently reliable direct estimates of the monthly unemployment rate. Our objective was to investigate whether the Fay-Herriot model could be used to obtain monthly estimates that would be reliable enough to be published.

We constructed an auxiliary variable z_{1i}, for area i, given by $z_{1i} = N_i^{\text{EIB}} / N_i^{15+}$, where N_i^{EIB} is the number of employment insurance beneficiaries in area i and N_i^{15+} is the number of persons aged 15 years or older in area i. The numerator is obtained from an administrative source, whereas the denominator is a Census projection computed by Statistics Canada. We used the vector $\mathbf{z}_i = (1, z_{1i})^T$, along with $b_i = 1$, $i = 1, \ldots, m$, to obtain SAE estimates. We used May 2016 data in this investigation to allow the comparison of direct and SAE estimates with 2016 Census estimates.

Some of the 149 areas of interest had a very small sample size in the LFS: they were not used in the Fay-Herriot and smoothing models. As a rule of thumb, we excluded from the models, areas where the number of sampled persons in the labour force was smaller than 10. There were 9 such areas; among them, six had no sampled person in the labour force. Also, there were 9 other areas where the direct unemployment rate estimate, $\hat{\theta}_i$, and its direct variance estimate, $\hat{\psi}_i$, were both equal to 0. As these direct estimates were not deemed to be reliable enough, their associated areas were excluded from the models. This resulted in using only 131 areas in the models. For those areas, the small area estimates are EBLUP estimates, with the remaining 18 being synthetic estimates.

The estimator $\hat{\psi}_i$ of the direct variance ψ_i was obtained via the Rao-Wu bootstrap. The estimates of the smooth design variances were then obtained by using $\mathbf{x}_i = (1, \log(z_{1i}), \log(1 - z_{1i}), \log(N_i^{15+}))^T$. A graph of the residuals of the smoothing model, $\log(\hat{\psi}_i) - \mathbf{x}_i'\hat{\boldsymbol{\alpha}}$, versus the predicted values, $\mathbf{x}_i'\hat{\boldsymbol{\alpha}}$, did not reveal any obvious model misspecification. Figure 6.1 shows a graph of direct variances estimates, $\hat{\psi}_i$, versus smooth variance estimates, $\hat{\tilde{\psi}}_i$. The red line is the identity line. If the smoothing model is appropriate, for any value of $\hat{\tilde{\psi}}_i$, the average of direct variance estimates for areas around area i should be roughly equal to $\hat{\tilde{\psi}}_i$. This means that the red line should pass roughly through the middle of the points everywhere. From a quick inspection of Figure 6.1, we observe that the red line is close to the middle of the points although probably slightly above the middle due to some extreme values of $\hat{\psi}_i$. This may result in a slight overestimation of the true smooth variance $\tilde{\psi}_i = E_{mp}(\hat{\psi}_i)$. A slight overestimation is not a major issue. What has to be avoided is an underestimation of $\tilde{\psi}_i$, as it typically leads to underestimating the MSE of the SAE estimate. This would provide the user with a false impression of precision.

Overall, we were satisfied with our smoothed variance estimates. However, for areas with large sample sizes, we set $\hat{\tilde{\psi}}_i = \hat{\psi}_i$ as our estimate of $\tilde{\psi}_i$. We assumed that direct variance estimates were stable enough when the sample size is large. As a rule of thumb, we set $\hat{\tilde{\psi}}_i = \hat{\psi}_i$ when the number of sampled persons in the labour force was greater than 400. This replacement occurred for 35 areas. The strategy was used to avoid possible small model biases in $\hat{\tilde{\psi}}_i$ for the largest areas, which could result in EBLUP estimates that become significantly different from the direct estimates. This is not a desirable property for areas with a large sample size.

The smooth variance estimates were then used to obtain small area estimates for the 149 areas of interest. Figure 6.2 shows a graph of small area and direct estimates as a function of sample size (number of sampled persons in the labour force). The small area estimates are much less volatile than direct estimates, especially for the areas with the smallest sample sizes. For the largest areas, as expected, both estimates are similar.

We first evaluated the quality of the underlying Fay-Herriot model before looking at the MSE estimates. Figure 6.3 shows the graph of direct estimates, $\hat{\theta}_i$, versus predicted values, $\mathbf{z}_i^T\hat{\boldsymbol{\beta}}$. The red line is the identity line and the blue line is a nonparametric smoothing spline curve. If the linearity assumption holds, the blue line should be close to the red line and the latter should pass roughly through the middle of the points everywhere. Figure 6.3 does not give any indication that the linearity assumption of the Fay-Herriot model is questionable.

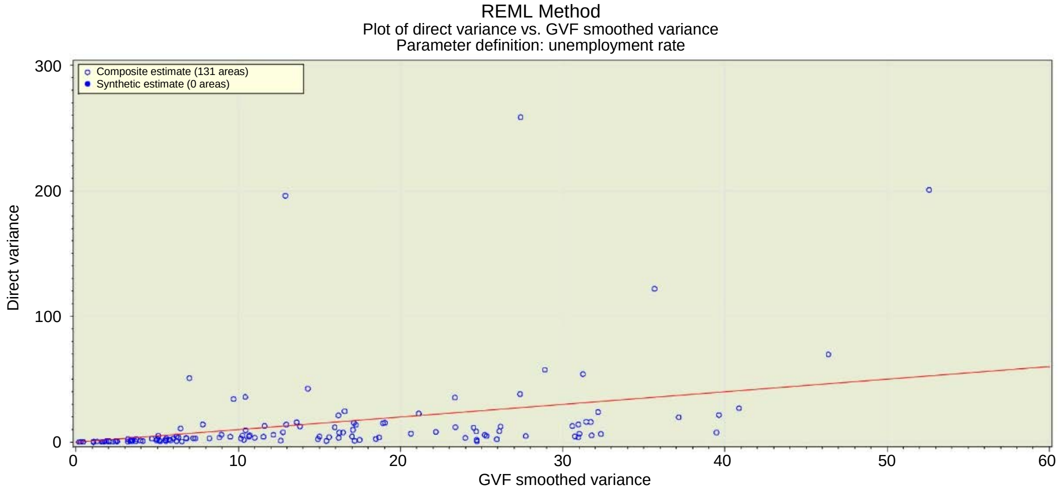

Figure 6.1 Graph of direct variance estimates, $\hat{\psi}_i$, versus smooth variance estimates, $\hat{\tilde{\psi}}_i$.

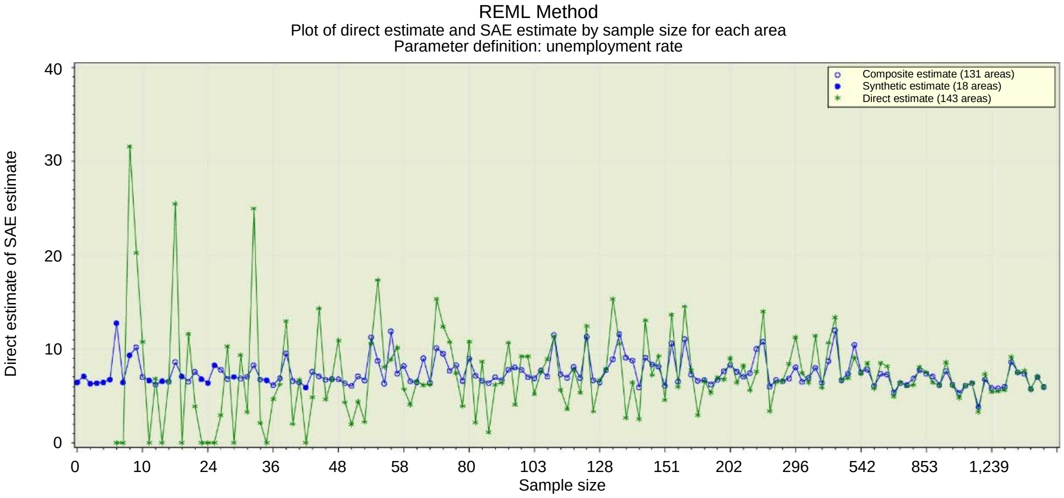

Figure 6.2 Graph of small area estimates and direct estimates as a function of sample size.

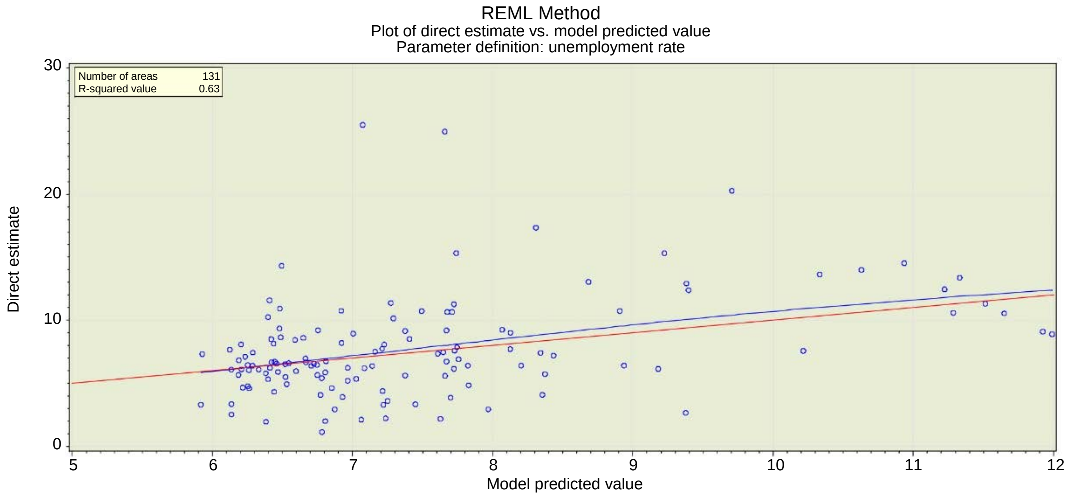

Figure 6.3 Graph of direct estimates versus model predicted values.

It is also informative to compute a measure that indicates the strength of $\mathbf{z}_i$ for the prediction of θ_i. To this end, we developed and implemented a coefficient of determination, or R^2 value, associated with the linking model $\theta_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i$. Note that the coefficient of determination associated with the combined model, $\hat{\theta}_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i + e_i$, is not of interest as the objective is not the prediction of $\hat{\theta}_i$ but the prediction of θ_i. Our coefficient of determination is given by

$$R^2 = 1 - \frac{\hat{\sigma}_v^2}{\frac{(m-q)}{(m-1)} \hat{\sigma}_v^2 + S^2\left(\hat{\boldsymbol{\beta}}\right)},$$

where q is the dimension of $\mathbf{z}_i$ and $S^2\left(\hat{\boldsymbol{\beta}}\right)$ is the sample variance of $\mathbf{z}_i^T \hat{\boldsymbol{\beta}} \big/ b_i$ (see equation (A.6) for the exact definition of the function $S^2(\cdot)$). The details of the derivation of the above coefficient of determination are provided in the Appendix. Figure 6.3 indicates that the R^2 value is 0.63. The linking model is thus neither weak nor extremely strong but, hopefully, strong enough to achieve efficiency gains over the direct estimator. The system also produces estimates of the parameters of the Fay-Herriot model along with their standard errors. From this output, we found out that estimates of both the intercept and slope parameters of the Fay-Herriot model were significantly different from 0 using a standard Wald test at the 0.05 significance level.

Figure 6.4 shows a graph of standardized residuals, $\left(\hat{\theta}_i - \mathbf{z}_i^T \hat{\boldsymbol{\beta}}\right) \Big/ \sqrt{b_i^2 \hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i}$, versus standardized predicted values, $\mathbf{z}_i^T \hat{\boldsymbol{\beta}} \Big/ \sqrt{b_i^2 \hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i}$. The red line is a horizontal line at zero and the blue line is a nonparametric smoothing spline curve. Similarly to Figure 6.3, the blue line should be close to the red line under linearity and the latter should pass roughly through the middle of the points everywhere. Again, Figure 6.4 does not indicate any obvious failure of the linearity assumption underlying the Fay-Herriot model.

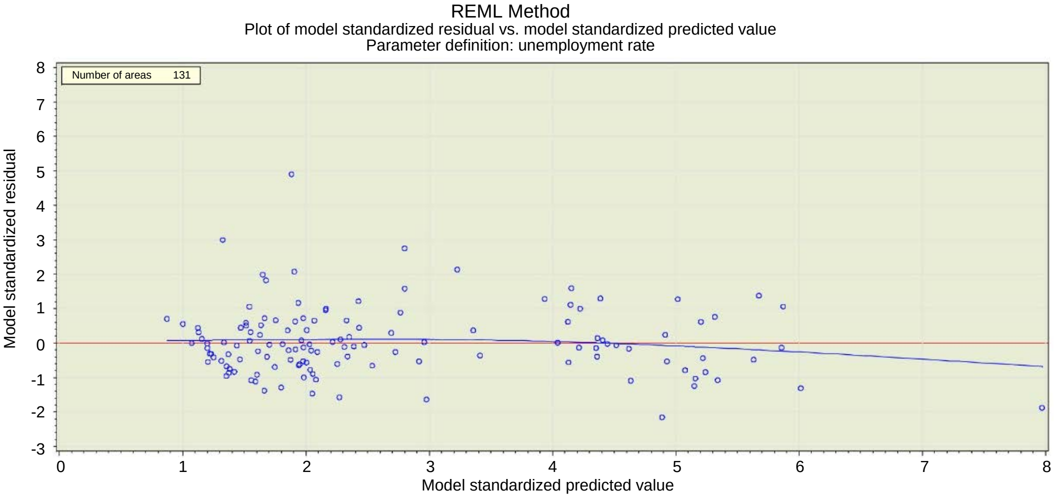

Figure 6.4 Graph of standardized residuals versus standardized predicted values.

Figure 6.5 shows a graph of squared standardized residuals versus standardized predicted values. The red line is a horizontal line at one and the blue line is again a nonparametric smoothing spline curve. This graph is used to check the homoscedasticity assumption; i.e., the assumption that the model variance σ_v^2 is constant. Under homoscedasticity, the blue line should be close to the red line everywhere. The graph does not reveal any obvious presence of heteroscedasticity.

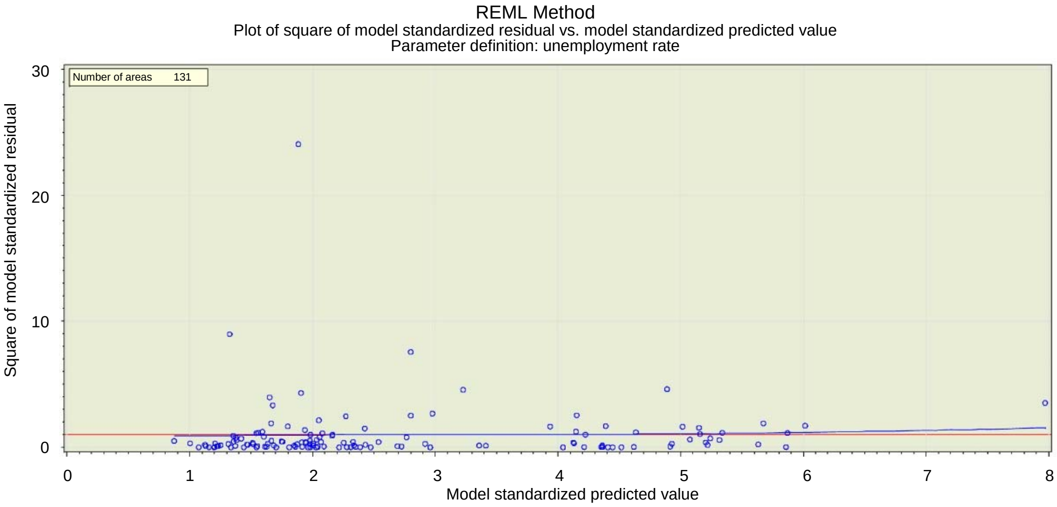

Figure 6.5 Graph of square standardized residuals versus standardized predicted values.

Figure 6.6 shows a QQ-plot of standardized residual quantiles versus standard normal quantiles. It is used to verify the normality assumption of the errors $b_i v_i$ and e_i. The graph does indicate a modest departure from normality. However, Rao and Molina (2015, page 138) argued that EBLUP estimates and their corresponding MSE estimates are generally robust to deviations from normality.

The system also computes Cook's distances to identify areas that could have a significant influence on the estimate $\hat{\boldsymbol{\beta}}$. The Cook distance for area i is given by

$$D_i = \frac{1}{q}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)}\right)^T \sum_{j=1}^{m} \frac{\mathbf{z}_j \mathbf{z}_j^T}{b_j^2 \hat{\sigma}_v^2 + \hat{\tilde{\psi}}_j} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)}\right),$$

where $\hat{\boldsymbol{\beta}}^{(-i)}$ is the estimate of $\boldsymbol{\beta}$ obtained after deleting area i. A plot of the influences D_i is provided in Figure 6.7. One area seems to have a relatively large influence compared with other areas $(D_i = 1.2851)$. This area has the largest standardized predicted value and the second largest predicted value. Its standardized residual is -1.88, which is not extreme, although not very small either. Its sample size is large (number of sampled persons in the labour force close to 500) and its smooth variance estimate, $\hat{\tilde{\psi}}_i$, is relatively small compared with other areas. All these reasons explain why this area was detected as being influential. In this application, we decided to keep this area in the model as its influence was not large enough to make a big difference in the SAE estimates and their corresponding MSE estimates.

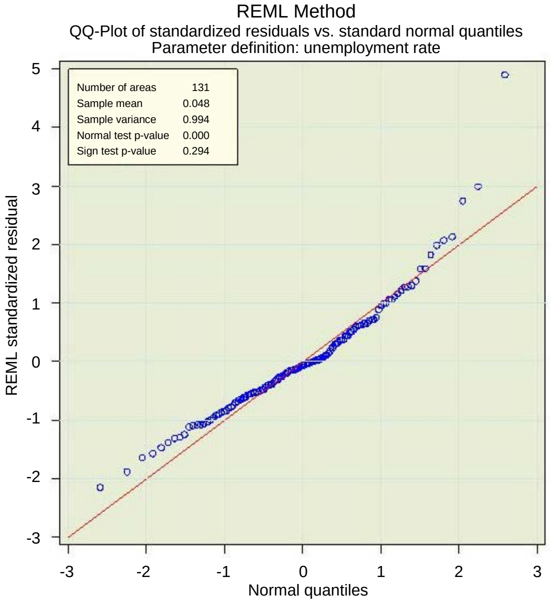

Figure 6.6 QQ-plot of standardized residual quantiles versus standard normal quantiles.

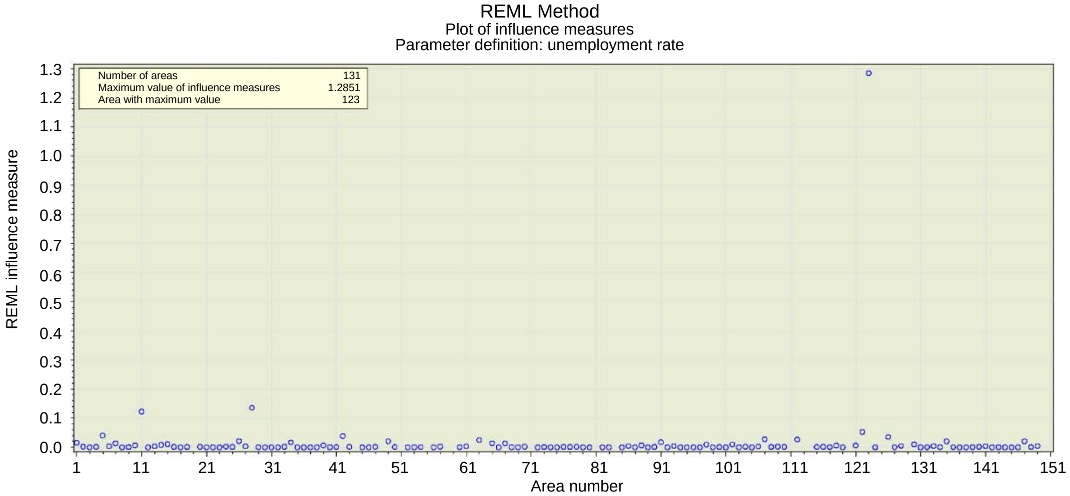

Figure 6.7 Plot of Cook's distances.

Since the Fay-Herriot model and smoothing model were both reasonable, we computed MSE estimates to evaluate the magnitude of the efficiency gains, if any, obtained by using the Fay-Herriot model. Figure 6.8 shows the estimated direct Coefficient of Variation (CV), defined as $\sqrt{\hat{\hat{\psi}}_i} \Big/ \hat{\theta}_i$, and the estimated SAE Relative Root Mean Square Error (RRMSE), defined as $\sqrt{\hat{\phi}_i} \Big/ \hat{\theta}_i^{\text{SAE}}$, where $\hat{\phi}_i$ is an estimate of the MSE, $E_{mp}\left(\hat{\theta}_i^{\text{SAE}} - \theta_i\right)^2$, and $\hat{\theta}_i^{\text{SAE}}$ is the small area estimate (EBLUP or synthetic estimate) of θ_i. The sample size (number of sampled persons in the labour force) is given on the horizontal axis. The estimated direct CVs are in general much larger than the estimated SAE RRMSEs, especially for the areas with the smallest sample sizes. The estimated SAE RRMSEs are never above 20% whereas the estimated direct CV is over

300% for one area. The estimated SAE RRMSEs are also very stable as a function of the sample size unlike the erratic behavior of the estimated direct CVs. For the areas with the largest sample sizes, both estimates are very similar, as expected. This indicates that SAE methods can lead to a substantial increase of precision over direct estimation methods, particularly for the smallest areas.

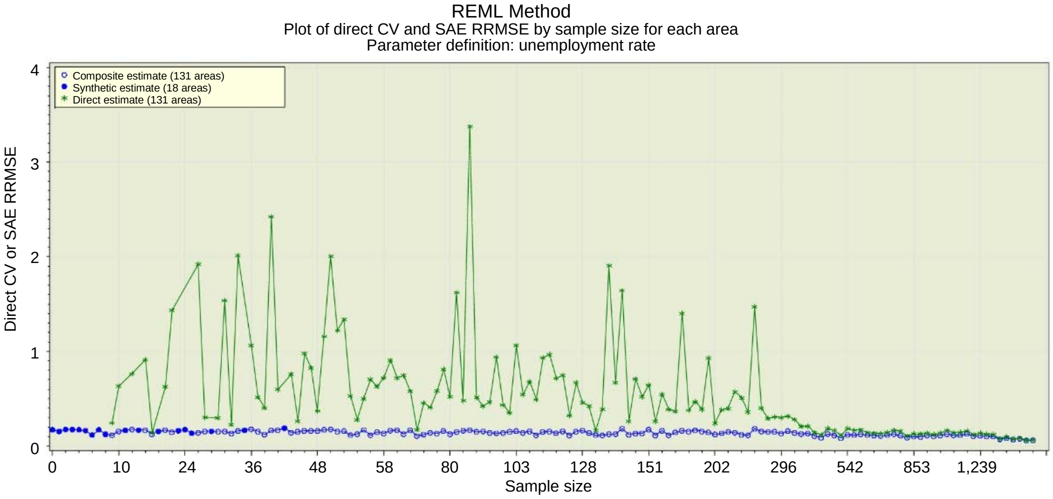

Figure 6.8 Graph of estimated direct CVs and SAE RRMSEs as a function of sample size.

For the month of May 2016, we had the luxury of having a very reliable source for the estimation of the unemployment rates: the 2016 long form Census administered to roughly one-fourth of the households throughout Canada. The Census sample size is much larger than the LFS sample size in all the areas of interest. Therefore, we used the 2016 Census direct estimates, denoted by $\hat{\theta}_i^{\text{Census}}$, as a gold standard for evaluating the accuracy of both the LFS direct estimates and SAE estimates. We computed Absolute Relative Differences (ARD) between LFS direct estimates and Census estimates, $\left|\hat{\theta}_i - \hat{\theta}_i^{\text{Census}}\right| \big/ \hat{\theta}_i^{\text{Census}}$, as well as ARDs between SAE estimates and Census estimates, $\left|\hat{\theta}_i^{\text{SAE}} - \hat{\theta}_i^{\text{Census}}\right| \big/ \hat{\theta}_i^{\text{Census}}$. These ARDs were then averaged within 5 different homogeneous subgroups with respect to sample size. Table 6.1 summarizes the results.

Table 6.1
Average ARD of SAE estimates and LFS direct estimates expressed in percentage

Sample size	Average ARD between LFS direct estimates and Census estimates	Average ARD between SAE estimates and Census estimates	Average ARD between HB estimates and Census estimates
28 smallest areas	70.4%	17.7%	18.3%
Next 28 smallest areas	38.7%	18.9%	19.0%
Next 28 smallest areas	26.2%	13.8%	14.1%
Next 28 smallest areas	20.9%	12.7%	13.0%
28 largest areas	13.2%	10.2%	10.3%
Overall	33.9%	14.7%	14.9%

Note: Out of the 149 areas of interest in the LFS, 9 were excluded from this table: six where the LFS number of sampled persons in the labour force was 0 and three that were no longer in the list of CMAs /CAs after the 2016 Census.

As expected, the ARD between the LFS and Census direct estimates decreases as the sample size increases. This may suggest that the conceptual differences between these two surveys and nonsampling errors are reasonably small compared with the sampling error, especially for the smallest areas where the sampling error may be the main contributor to the ARD. The SAE estimates are much closer to the Census estimates than the LFS direct estimates, particularly for the smallest areas where improvement is most needed. This confirms that our underlying models are reasonable in this application.

For comparison purposes, we also computed HB estimates, $\hat{\theta}_i^{\text{HB}},$ based on the matched Fay-Herriot model with the noninformative priors for $\boldsymbol{\beta}$, σ_v^2 and $\tilde{\psi}_i$ provided in Section 5. We then computed ARDs between HB estimates and Census estimates, $\left|\hat{\theta}_i^{\text{HB}} - \hat{\theta}_i^{\text{Census}}\right| \big/ \hat{\theta}_i^{\text{Census}}.$ Results are given in the last column of Table 6.1. The averaged ARDs of the HB estimates are close to those of the EBLUP estimates.

7 Conclusion

A frequent demand from users of data from NSOs is for more granularity for use in planning and policy research purposes. NSOs can no longer simply increase the sample sizes of their surveys to obtain reliable estimates at the requested level of detail. Reasons for this include the high costs of doing so, response burden concerns, as well as the difficult task of obtaining responses from sampled units. An alternative being investigated by many NSOs is the use of small area estimation techniques that provide a way to address the demand for more granular data. With this in mind, Statistics Canada began the development of an SAE production system in the early 2000s and now have such a system available to their statistical programs. The production system handles area and unit level models, with multiple options such as different methods to estimate the variance components, different linking models and both the EBLUP and HB estimation methods. It is currently being used to produce experimental estimates for several Statistics Canada statistical programs and it is expected that the first published small area estimates will be available in 2019.

As it was mentioned in the introduction, the only existing software in 2006 that would produce small area estimates and their associated mean squared estimates was sponsored by the EURAREA (2004) project. The current production system developed at Statistics Canada is written in SAS, with its methodology closely following Rao (2003) and includes some recent advances. As it stands, it satisfies the existing requirements for small area estimation at Statistics Canada. However, as the use of small area estimation becomes more common within Statistics Canada, there will be a need to add functionality to the system to meet this demand. The recent book authored by Rao and Molina (2015) provides an idea of how much development has taken place in small area estimation during recent years. The incorporation of all this development into the production system would be extremely time consuming, expensive, and may not be directly applicable to the needs of Statistics Canada. It, therefore, follows that options other than programming these new functionalities in the current SAS production system should be considered. One option would be to investigate how packages developed elsewhere, such as those written in $R,$ can be integrated into it. Notable packages written in R include *sae* (Molina and Marhuenda, 2015), *mme*

(Lopez-Vizcaino, Lombardia and Morales, 2014), *saery* (Esteban, Morales and Perez, 2014) and *sae2* (Fay and Diallo, 2015). These packages include small area procedures that are not in the present system such as multinomial linear mixed models, area level models with time effects and time series area level models supporting univariate and multivariate applications. The existing SAS production system meets the needs of Statistics Canada at this point in time, and there are no concrete plans to add functionality to it.

Acknowledgements

We would like to thank the reviewers for their comments and suggestions that led to improvements in this paper.

Appendix

Justification of the coefficient of determination

In order to determine a coefficient of determination associated with the linking model, $\theta_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i$, we first rewrite it as

$$\tilde{\theta}_i = \tilde{\mathbf{z}}_i^T \boldsymbol{\beta} + v_i,$$

where $\tilde{\theta}_i = \theta_i / b_i$ and $\tilde{\mathbf{z}}_i = \mathbf{z}_i / b_i$. We assume that an intercept is implicitly or explicitly included in $\tilde{\mathbf{z}}_i$; i.e., there exists a vector $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^T \tilde{\mathbf{z}}_i = 1$. In other words, we assume that there exists a vector $\boldsymbol{\lambda}$ such that $b_i = \boldsymbol{\lambda}^T \mathbf{z}_i$. If $\tilde{\theta}_i$, $i = 1, \ldots, m$, were known, we could estimate the unknown vector of model parameters $\boldsymbol{\beta}$ by the least squares estimator

$$\hat{\boldsymbol{\beta}}_* = \left(\sum_{i=1}^m \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right)^{-1} \sum_{i=1}^m \tilde{\mathbf{z}}_i \tilde{\theta}_i$$

and the unknown model variance σ_v^2 by the unbiased estimator

$$\hat{\sigma}_{v*}^2 = \frac{\sum_{i=1}^m \left(\tilde{\theta}_i - \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* \right)^2}{m - q}.$$

The well-known adjusted coefficient of determination is

$$R_{\text{ideal}}^2 = 1 - \frac{\hat{\sigma}_{v*}^2}{(m-1)^{-1} \sum_{i=1}^m \left(\tilde{\theta}_i - \bar{\tilde{\theta}} \right)^2}, \tag{A.1}$$

where $\bar{\tilde{\theta}} = m^{-1} \sum_{i=1}^m \tilde{\theta}_i$. It is an ideal coefficient of determination because it cannot be computed (since $\tilde{\theta}_i$ is unknown) but this is the target we would like to estimate. Simply replacing θ_i with $\hat{\theta}_i$ does not solve the problem as $\hat{\theta}_i$ reflects the combined model and not just the linking model. The resulting coefficient of determination would typically be too small. To obtain a better estimate of R^2_{ideal}, we first decompose $\sum_{i=1}^m \left(\tilde{\theta}_i - \bar{\tilde{\theta}}\right)^2$ as

$$\sum_{i=1}^m \left(\tilde{\theta}_i - \bar{\tilde{\theta}}\right)^2 = \sum_{i=1}^m \left(\tilde{\theta}_i - \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_*\right)^2 + \sum_{i=1}^m \left(\tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* - \bar{\tilde{\theta}}\right)^2 + 2\sum_{i=1}^m \left(\tilde{\theta}_i - \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_*\right)\left(\tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* - \bar{\tilde{\theta}}\right). \tag{A.2}$$

Assuming that an intercept is implicitly or explicitly included in $\tilde{\mathbf{z}}_i$ and from the expression for $\hat{\boldsymbol{\beta}}_*$, we have that

$$\sum_{i=1}^m \left(\tilde{\theta}_i - \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_*\right) \tilde{\mathbf{z}}_i = \mathbf{0} \tag{A.3}$$

and

$$\sum_{i=1}^m \left(\tilde{\theta}_i - \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_*\right) = 0. \tag{A.4}$$

From (A.4), we can rewrite $\bar{\tilde{\theta}}$ as $\bar{\tilde{\theta}} = \bar{\tilde{\mathbf{z}}}^T \hat{\boldsymbol{\beta}}_*$, where $\bar{\tilde{\mathbf{z}}} = m^{-1} \sum_{i=1}^m \tilde{\mathbf{z}}_i$. As a result, the cross product term in (A.2) vanishes and equation (A.2) reduces to

$$\begin{aligned} \sum_{i=1}^m \left(\tilde{\theta}_i - \bar{\tilde{\theta}}\right)^2 &= \sum_{i=1}^m \left(\tilde{\theta}_i - \tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_*\right)^2 + \sum_{i=1}^m \left(\tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* - \bar{\tilde{\mathbf{z}}}^T \hat{\boldsymbol{\beta}}_*\right)^2 \\ &= (m-q)\,\hat{\sigma}^2_{v*} + (m-1)\,S^2\left(\hat{\boldsymbol{\beta}}_*\right), \end{aligned} \tag{A.5}$$

where

$$S^2\left(\hat{\boldsymbol{\beta}}_*\right) = \frac{\sum_{i=1}^m \left(\tilde{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}_* - \bar{\tilde{\mathbf{z}}}^T \hat{\boldsymbol{\beta}}_*\right)^2}{m-1}. \tag{A.6}$$

From (A.5), it follows that the ideal coefficient of determination (A.1) can be rewritten as

$$R^2_{\text{ideal}} = 1 - \frac{\hat{\sigma}^2_{v*}}{\frac{(m-q)}{(m-1)}\hat{\sigma}^2_{v*} + S^2\left(\hat{\boldsymbol{\beta}}_*\right)} \equiv f\left(\hat{\boldsymbol{\beta}}_*, \hat{\sigma}^2_{v*}\right). \tag{A.7}$$

The only unknown quantities in (A.7) are $\hat{\boldsymbol{\beta}}_*$ and $\hat{\sigma}^2_{v*}$. A computable coefficient of determination can thus be obtained by replacing $\hat{\boldsymbol{\beta}}_*$ and $\hat{\sigma}^2_{v*}$ in (A.7) with $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2_v$, the consistent estimators of $\boldsymbol{\beta}$ and σ^2_v implemented in the SAE system and described in Section 3. The resulting coefficient of determination can be expressed as $R^2 = f\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_v\right)$, with the function $f(\cdot,\cdot)$ defined in (A.7), and is a consistent estimator of the ideal coefficient of determination R^2_{ideal}.

References

Battese, G.E., Harter, R.M. and 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.

Beaumont, J.-F., and Bocci, C. (2016). Small area estimation in the Labour Force Survey. Paper presented at the Advisory Committee on Statistical Methods, May 2016, Statistics Canada.

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

Chib, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician*, 49, 327-335.

Dick, P. (1995). Modelling net undercoverage in the 1991 Canadian Census. *Survey Methodology*, 21, 1, 45-54. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1995001/article/14411-eng.pdf.

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, 1, 17-47. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1982001/article/14328-eng.pdf.

Esteban, M.D., Morales, D. and Perez, A. (2014). saery: Small Area Estimation for Rao and Yu Model. URL http://CRAN.R-project.org/package=saery. R package version 1.0.

Estevao, V., Hidiroglou, M.A. and You, Y. (2015). *Area Level Model, Unit Level, and Hierarchical Bayes Methodology Specifications*. Internal document, Statistics Canada.

EURAREA (2004). *Enhancing Small Area Estimation Techniques to meet European Needs*. https://cordis.europa.eu/project/rcn/58374_en.html.

Fay, R.E., and Diallo, M. (2015). sae2: Small Area Estimation: Time-Series Models. URL http://CRAN.Rproject.org/package=sae2. R package version 0.1-1.

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., and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 1, 45-51. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001001/article/5853-eng.pdf.

Gambino, J., Kennedy, B. and Singh, M.P. (2001). Regression composite estimation for the Canadian Labour Force Survey: Evaluation and implementation. *Survey Methodology*, 27, 1, 65-74. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2001001/article/5855-eng.pdf.

Gelfand, A.E., and Smith, A.F.M. (1990). Sample-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 972-985.

Ghangurde, P.D., and Singh, M.P. (1977). Synthetic estimation in periodic household surveys. *Survey Methodology*, 3, 2, 152-181.

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.

Kott, P. (1989). Robust small domain estimation using random effects modeling. *Survey Methodology*, 15, 1, 3-12. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1989001/article/14581-eng.pdf.

Li, H., and Lahiri, P. (2010). Adjusted maximum method in the small area estimation problem. *Journal of Multivariate Analysis*, 101, 882-892.

Lopez-Vizcaino, E., Lombardia, M.J. and Morales, D. (2014). mme: Multinomial Mixed Effects Models, 2014. URL http://CRAN.R-project.org/package=mme. R package version 0.1-5.

Molina, I., and Marhuenda, Y. (2015). sae: An R package for small area estimation. *The R Journal*, 7, 1, 81-98.

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, 1, 67-72. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1999001/article/4713-eng.pdf.

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

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rivest, L.-P., and Belmonte, E. (2000). A conditional mean squared error of small area estimators. *Survey Methodology*, 26, 1, 67-78. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2000001/article/5179-eng.pdf.

Rubin-Bleuer, S. (2014). *Specifications for EBLUP and Pseudo-EBLUP Estimators with Nonnegligible Sampling Fractions*. Statistics Canada document.

Rubin-Bleuer, S., Jang, L. and Godbout, S. (2016). The Pseudo-EBLUP estimator for a weighted average with an application to the Canadian Survey of Employment, Payrolls and Hours. *Journal of Survey Statistics and Methodology*, 4, 417-435.

Singh, M.P., and Tessier, R. (1976). Some estimators for domain totals. *Journal of the American Statistical Association*, 71, 322-325.

Singh, A.C., Kennedy, B. and Wu, S. (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating panel design. *Survey Methodology*, 27, 1, 33-44. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001001/article/5852-eng.pdf.

Stukel, D., and Rao, J.N.K. (1997). Small-area estimation under two-fold nested error regression model. *Journal of Statistical Planning and Inference*, 78, 131-147.

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

Wang, J., Fuller, W.A. and Qu, Y. (2008). Small area estimation under a restriction. *Survey Methodology*, 34, 1, 29-36. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2008001/article/10619-eng.pdf.

You, Y., and Rao, J.N.K. (2002). A pseudo empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431-439.

You, Y., Rao, J.N.K. and Dick, P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.

You, Y., Rao, J.N.K. and Hidiroglou, M. (2013). On the performance of self benchmarked small area estimators under the Fay-Herriot area level model. *Survey Methodology*, 39, 1, 217-229. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11830-eng.pdf.