Combining Survey Data with Other Data Sources Using Small Area Estimation: A Case Study

Golshid Chatrchi[1], Héloïse Gauvin[1], Allen Ding[1]
[1]Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa ON K1A 0T6

Abstract
With the increasing availability of large data sources, there is an interest in combining survey data with information from these other sources to improve the quality of domain-level statistics. An area-level model approach is considered in the context of small area estimation to integrate survey data with an aggregated source of information. The proposed method is applied to the real problem of estimating the inbound tourism spending, using aggregated Payment Processors' data and survey data from Visitor Travel Survey (VTS) in Canada. The area-level model is also used to provide timely forecast estimates of foreign tourism spending, using current aggregated Payment Processors' data and historical VTS data.

Key Words: Small Area Estimation, Data Integration, Modeling, Area-level Model, Prediction, Small domains

1. Introduction

The demand for domain-level estimates has been increasing over recent years. Surveys conducted by national statistical agencies produce reliable estimates for domains with sufficient sample units. However, the survey estimates may be unreliable when the domain of interest has few sample units. Moreover, with increasing levels of nonresponse in household surveys, producing data at finer levels of detail has become more challenging. To reduce response burden and improve standard (direct) estimates derived from surveys, Statistics Canada has been investigating alternative data sources. Alternative data sources may include administrative data and other source of information that are not collected from surveys. One possible solution is to combine survey data and alternative data source using Small Area Estimation (SAE) methods. In this way, we can improve the quality of domain-level statistics and benefit from the positive features of survey data and other available data sources.

In this paper we focus on the Visitor Travel Survey (VTS) in Canada and show through an example that combining survey data with alternative data sources can improve domain-level estimates. Section 2 gives a brief description of the SAE methodology with a focus on the area-level model and smoothing design variances. The Visitor Travel Survey, which is the main focus of this study, and the alternative data source used in the model, Payment data, are discussed in Section 3 and Section 4 respectively. In Section 5 and Section 6, we review the application to VTS data, the modelling steps and present some results. We summarize our discussion in Section 7.

2. Small Area Estimation

The idea behind Small Area Estimation (SAE) is to produce reliable estimates for small domains, where the standard direct estimates calculated using weighted survey responses (such as the Narain-Horwitz-Thompson estimator) cannot be used due to unacceptably large standard errors for areas with small sample sizes. In such situations, it is necessary to use indirect estimators that "borrow strength" from related areas. These indirect estimators tend to be more efficient than direct estimators since they increase the effective sample size by incorporating a number of small areas in a single model. The small area estimate has two parts: the direct estimate from the survey data, and a prediction based on a model.

The SAE models can be classified into two types: area-level (or aggregate) models that relate small area means to area-specific covariates, and unit-level models that relate the unit values of the study variable to unit-specific covariates. Rao and Molina (2015) is a comprehensive reference in this area.

In this study, we apply the area-level model as the auxiliary information used in the model is aggregated. The theory of the area-level model is briefly described in the following subsection.

2.1 Area-level model

A basic area-level model, known as Fay-Herriot model (Fay and Herriot, 1979) has two parts:

1. Sampling model: $\hat{\theta}_i = \theta_i + e_i$,

where $\hat{\theta}_i$ is the direct estimator of variable of interest for area *i*, e_i represents the sampling errors with $E_p(e_i) = 0$ and $\text{var}_p(e_i) = \psi_i$. The subscript *p* indicates that the expectation and variance are taken with respect to the sampling design (or the sample selection mechanism). The implicit assumption is that the direct estimator is unbiased under the sampling design. The quantity ψ_i represents the variance of $\hat{\theta}_i$ with respect to the sampling design and is typically unknown. A direct estimator of ψ_i is denoted by $\hat{\psi}_i$

2. Linking model: $\theta_i = \boldsymbol{z}_i{}^T\boldsymbol{\beta} + b_i v_i$

The vector of $\boldsymbol{z}_i$ represents the area specific auxiliary data, v_i is the area-specific random effects with $E_m(v_i|\boldsymbol{z}_i) = 0$ and $\text{var}_m(v_i|\boldsymbol{z}_i) = \sigma_v^2$, b_i is a positive constant and $\boldsymbol{\beta}$ and σ_v^2 are unknown model parameters. The subscript *m* indicates that the expectation and variance are taken with respect to the model.

In addition to the above model assumptions, the errors e_i and v_i, *i=1,..., M*, are usually assumed to be normally distributed and mutually independent. The quantity *M* is the number of areas used for modelling. By combining the sampling and linking model, we obtain the Fay-Herriot model:

$\hat{\theta}_i = \boldsymbol{z}_i{}^T\boldsymbol{\beta} + a_i$,

where $a_i = b_i v_i + e_i$, $E_{mp}(a_i|\boldsymbol{z}_i) = 0$, $var_{mp}(a_i|\boldsymbol{z}_i) = b_i^2\sigma_v^2 + \tilde{\psi}_i$ and $\tilde{\psi}_i = E_m(\psi_i|\boldsymbol{z}_i)$ is a smoothed design variance[1]. The subscript *mp* indicates that the expectation and variance are taken with respect to both the model and sampling design.

[1] Estimation of the smoothed design variance is discussed in Sub-section 2.2.

Assuming that $\tilde{\psi}_i$ and σ_v^2 are known, the theory of general linear mixed models provides a framework to derive the optimal predictor for θ_i called Best Linear Unbiased Predictor or BLUP.
The BLUP can be expressed as the weighted combination of direct estimator ($\hat{\theta}_i$) and "regression synthetic" estimator ($\boldsymbol{z}_i{}^T\tilde{\boldsymbol{\beta}}$):
$\tilde{\theta}_i = \gamma_i\hat{\theta}_i + (1-\gamma_i)\boldsymbol{z}_i{}^T\tilde{\boldsymbol{\beta}}$,
where $\tilde{\boldsymbol{\beta}} = \left[\sum_{i=1}^m \boldsymbol{z}_i\,\boldsymbol{z}_i{}^T/(b_i^2\sigma_v^2+\tilde{\psi}_i)\right]^{-1}\left[\sum_{i=1}^m \boldsymbol{z}_i\,\hat{\theta}_i/(b_i^2\sigma_v^2+\tilde{\psi}_i)\right]$ and $\gamma_i = \frac{b_i^2\sigma_v^2}{b_i^2\sigma_v^2+\tilde{\psi}_i}$.

In practice, $\tilde{\psi}_i$ and σ_v^2 are usually unknown, and are replaced by their estimators. The result is the Empirical Best Linear Unbiased Predictor (EBLUP):

$$\hat{\theta}_i^{EBLUP} = \hat{\gamma}_i\hat{\theta}_i + (1-\hat{\gamma}_i)\boldsymbol{z}_i{}^T\hat{\boldsymbol{\beta}},$$

where $\hat{\gamma}_i = b_i^2\hat{\sigma}_v^2/b_i^2\hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i$ and $\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^m \boldsymbol{z}_i\,\boldsymbol{z}_i{}^T/(b_i^2\hat{\sigma}_v^2+\hat{\tilde{\psi}}_i)\right]^{-1}\left[\sum_{i=1}^m \boldsymbol{z}_i\,\hat{\theta}_i/(b_i^2\hat{\sigma}_v^2+\hat{\tilde{\psi}}_i)\right]$.

There are different procedures for estimating σ_v^2, such as the restricted maximum likelihood (REML), Fay-Herriot procedure (FH) as outlined in Fay and Herriot (1979), the Adjusted Density Maximization (ADM) by Li and Lahiri (2010), and the method of fitting constants (Henderson's method) . The main difference between these methods is how σ_v^2 is computed, using an iterative scoring algorithm.

We used REML method, which is based on the following (iterative) scoring algorithm:

$\hat{\sigma}_v^{2(a+1)} = \hat{\sigma}_v^{2(a)} + \left[\frac{1}{2}tr(\boldsymbol{PBPB})\right]^{-1}\left[\frac{1}{2}\boldsymbol{y}^T\boldsymbol{PBPy} - \frac{1}{2}tr(\boldsymbol{PB})\right]$,
where $\boldsymbol{B} = diag(b_i^2)$ is a diagonal matrix, $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1}$, $\mathrm{V} = diag(b_i^2\hat{\sigma}_v^{2(a)} + \tilde{\psi}_i)$, $\boldsymbol{y}$ is a column matrix with entries equal to $\hat{\theta}_i$, and $\boldsymbol{Z}$ is a matrix with row entries equal to $\boldsymbol{z}_i{}^T$.

2.2 Smoothing direct variances

Since the design variance ψ_i is unknown the smoothed design variance $\tilde{\psi}_i$ cannot be calculated. Instead, it is assumed that a design-unbiased variance estimator $\hat{\psi}_i$ is available from which the smoothed variance can be calculated. Although $\hat{\psi}_i$ is unbiased, it can be unstable when the area sample size is small. To address this issue, the variance estimate $\hat{\psi}_i$ was modelled and its predicted value was used as the estimate of the smoothed variance. More precisely, we have:

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

A simple unbiased estimator of the smooth design variance $\tilde{\psi}_i$ is $\hat{\psi}_i$. However, the latter may be quite inefficient when the domain sample size is small. A more stable estimator is obtained by modelling $\hat{\psi}_i$ given $\boldsymbol{z}_i$. Dick (1995), Rivest and Belmonte (2000) and Beaumont and Bocci (2016) used the method of Generalized Variance Functions (GVF). In our application, the GVF method did not provide satisfactory results, so it was decided to use a piecewise smoothing approach, instead. More details about the piecewise smoothing is given in Section 5.

For more information on estimation of the smoothed design variance see Hidiroglou, Beaumont and Yung (2019).

3. VTS data

The objective of the VTS is to provide a full range of statistics on the volume of international visitors to Canada and detailed characteristics of their trips such as expenditures, activities, places visited and length of stay. The target population of the VTS is all U.S. and overseas residents entering Canada. Excluded from the survey's coverage are diplomats and their dependents, refugees, landed immigrants, military, crew and former Canadian residents.

On a quarterly basis, the VTS provides information on the purpose of trip, size of travelling party, places visited, activities participated in during the trip, length of trip and trip spending. VTS data are collected in two ways: (i) border services officers distribute invitation cards at selected border points (the cards invite travellers to complete the electronic version of the questionnaire online); (ii) Statistics Canada interviewers visit selected international airports and interview international travellers in departure lounges with tablets. Airport interviews are based on sampling of predetermined time periods or flights. This part of the VTS is also referred to the Air Exit component. The primary objective of the Air Exit component is to improve the quality and reliability of trip and spending estimates for foreign air travellers to Canada, from major markets. The Air Exit component targets U.S. and overseas travellers at major airports across Canada returning directly to the USA or to selected overseas countries.
Invitation cards for the e-questionnaire are distributed at 137 designated ports of entry, and are actively distributed to U.S. and international travellers who enter Canada by one of the following modes of transportation: automobile, commercial plane, commercial bus or commercial boat.

In this study, we focus on the trip spending, which has six categories: transportation, accommodation, food and beverages, recreation, clothing and gift, and "other" spending. Direct domain-level estimates of tourism foreign spending can be obtained from the VTS, but they may not be reliable due to the small sample sizes.

4. Payment data

Statistics Canada received Payment processors' data from two data providers, through Destination Canada. The Payment data includes a portion of aggregated credit and debit card payments made by international visitors to Canada. The data is aggregated by Merchant Category Code[2] (MCC), FSA[3] (Forward Sorting Area) and the country of origin of the card.

Payment data has some limitations in terms of under-coverage or over-coverage. Cash spending or other payment methods are not covered in Payment data. Moreover, the Payment data consists of the aggregated spending information from only two data providers. Hence, the entire payment market is not covered and which creates an under-coverage issue. The other challenge, in terms of under-coverage, is related to travel booking websites. Travel packages that are bought on non-Canadian sites (i.e., domain extension not .ca) are not included in the Payment data.

[2] MCC is a four-digit number that classifies a business by the type of goods or services it provides.
[3] FSA code denotes a particular postal district and shows the alleged location of the business.

In terms of over-coverage, the Payment data includes some non-tourism spending such as, the inclusion of sales in merchant categories not related to tourism or expenditure made by non-travellers using a foreign credit or debit card. For example, spending by landed immigrants, who are still using their cards from back home, can be a part of Payment data. Also, Payment data could include purchases made on-line by someone who did not travel to Canada (i.e., exports) and thus could over-estimate tourism spending.

In addition to the coverage issues mentioned above, there are some inconsistencies between the targeted concept in the VTS and Payment data. In the VTS, for estimating (inbound) tourism spending, the targeted concept is spending made by international travellers related to trips that ended in a given quarter. However, in Payment data, the recorded information refers to the time of transactions which are not necessarily made during a trip but are made in advance of a trip. For example, spending related to accommodation or transportation are usually made ahead of the actual trip.

Despite these limitations, the correlation between payment data and VTS is fairly strong. The correlation coefficients are between 0.7 and 0.9, depending on the variable of interest.

5. Application to VTS data

As mentioned earlier, the goal is to investigate whether applying an area-level model to the data can improve the quality of VTS domain estimates, when an aggregated source of information is used as the auxiliary variable in the model. The domain of interest is defined based on the country of origin of visitors, and Tourism Regions (TR), which are geographical areas with certain boundaries in each of the provinces or territories. There are 84 TRs in 10 provinces and 3 territories in Canada. We tested the model and checked the availability of data and defined the domain of interest as 11 country groups by 22 groups of Tourism Regions. A separate model was fitted for each spending category. The domain of interest is defined as: 11 country groups × 22 grouped tourism regions (M=242).

The modelling steps can be classified into four main steps: 1) calculating direct estimates and their variances at the domain of interest, 2) smoothing direct variances, 3) fitting the SAE model and 4) model validation. The last two steps are repeated until satisfactory results are achieved.

The small area estimates are obtained through the use of the small area estimation module of the Statistics Canada's generalized software G-EST version 2.03 (Estevao et al., 2019).

5.1 Calculating VTS estimates and their variances

Let θ_i be the population parameter (i.e., inbound tourism spending) for a given domain i (and for a given quarter). The $\hat{\theta}_i$ values are calculated using survey weights:

$$\hat{\theta}_i = \sum_{k \in s_i} w_k y_k,$$

where y_k represents spending by unit k in domain i, and w_k is the sampling weight assigned to unit k on the VTS sample.

The variance of VTS estimates, $\hat{\psi}_i$, is calculated using mean bootstrap weights (Yung, 1997).

$$\hat{\psi}_i = \frac{R}{B} \sum_{b=1}^{B} (\theta_{ib} - \bar{\theta}_i)^2,$$

where B, the number of bootstrap replicates, R, the number of bootstrap samples, θ_{ib} is the estimate of total spending for domain "i" obtained from the b[th] bootstrap replicate and $\bar{\theta}_i$ is the mean of the totals obtained from B replicates.

5.2 Smoothing VTS variances

As mentioned earlier, smoothing direct variances is done to reduce the variability of $\hat{\psi}_i$. The resulting variances generally have less variability and fewer outliers. This usually leads to a better fit of the small-area models. The GVF method described in Section 2.2 didn't provide good results for VTS data. We used piecewise linear regression method (suggested by Beaumont and Bocci) for smoothing variances. The Piecewise linear regression is used to address possible nonlinear relationships between the dependent and independent variables. In a nutshell, the data points are divided in a certain number of consecutive segments. A linear function is assumed for each segment in such a way that the overall curve is continuous.

The estimator of the smooth design variance $\tilde{\psi}_i$ is obtained by applying a piecewise linear regression on the variance estimates $\hat{\psi}_i$:

$$\frac{\hat{\psi}_i}{\sqrt{X_i}} = \alpha_0 + \alpha_1 \frac{X_i}{\sqrt{X_i}} + \alpha_2 \frac{(X_i - c_1)^+}{\sqrt{X_i}} + \alpha_3 \frac{(X_i - c_2)^+}{\sqrt{Xi}} + \cdots + \alpha_{k+1} \frac{(X_i - c_k)^+}{\sqrt{Xi}} + \varepsilon_i,$$

where X_i is the spending from the Payment data and ε_i is a random error with zero mean and constant variance. The number of segments is chosen using a stepwise selection method with initial 9 cut-points. The square root transformation in the model is due to the structure of $\hat{\psi}_i$ (i.e., the relationship between $\hat{\psi}_i$ and X_i).

5.3 Piecewise area-level model

The basic area-level model did not provide satisfactory results, so we applied the piecewise area-level model, which is a modification of the basic area-level model. The piecewise area-level is useful when a single linear model does not provide an adequate explanation on the relationship between the variable of interest and covariates. The area specific auxiliary variable, X_i (i.e., spending from the Payment data), is partitioned into intervals and a separate line segment is fit to each interval:

$$\hat{\theta}_i^{SAE} = \beta_0 + \beta_1 X_i + \beta_2 (X_i - c_1)^+ + \beta_3 (X_i - c_2)^+ + \cdots + \beta_{k+1} (X_i - c_k)^+ + b_i v_i + e_i,$$

where c_k is the value of the k[th] breakpoint and $(X_i - c_k)^+ = \begin{cases} 0 & if\ X_i \le c_k \\ X_i - c_k & otherwise \end{cases}$ and b_i values are set equal to X_i . For the VTS data, the number of breakpoints is set to less than or equal to two and the cut-points are spaced equally.

It should be mentioned that there is no need to develop new theories for the piecewise area-level model, as the above equation can be easily written as the Fay-Herriot model:

$$\hat{\theta}_i^{SAE} = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) \boldsymbol{z}_i^T \hat{\boldsymbol{\beta}},$$

where $\boldsymbol{z}_i = (1, X_i, (X_i - c_1)^+, (X_i - c_2)^+, \ldots, (X_i - c_k)^+)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{k+1})$.

5.4 Model validation

The accuracy of small area estimates depends on the reliability of the Fay-Herriot model. It is therefore essential to make a careful assessment of the validity of the model before

releasing estimates. For instance, it is important to verify that a linear relationship actually holds between $\hat{\theta}_i$ and $\boldsymbol{z}_i$, at least approximately.

A simple way to verify the linearity assumption is to graph the standardized residuals, $\hat{a}_i$, against the predicted values $\boldsymbol{z}_i^T\hat{\boldsymbol{\beta}}$.

$$\hat{a}_i = \frac{\hat{\theta}_i - \boldsymbol{z}_i^T\hat{\boldsymbol{\beta}}}{\sqrt{b_i^2\hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i}}$$

The standardized residuals are key statistics that can also be used to verify other model assumptions such as the normality of the model errors. The linear assumption is reasonable when the graph does not reveal any particular trend. For the VTS data, the graph of standardized residuals vary by spending categories.

5.4.1 Outlier treatment

The outliers (i.e., areas that do not follow the same model as the other areas) of the area-level model need to be identified and if necessary the model should be re-fitted. Outliers are identified iteratively by examining the standardized residuals from that model. If $\hat{a}_i$'s are normal then $\hat{a}_i^2 \sim \mathcal{X}_1^2$. Let $\hat{a}_{im}^2$ be the largest squared standardized residual among the m domains used in the modelling. We find the value c such that $P(\hat{a}_{im}^2 \leq c) = 1 - \alpha$, for a given α. If the largest squared residual is greater than c (i.e.,$\hat{a}_{im}^2 > c$) then the corresponding domain is considered an outlier. That domain is set aside and the direct estimate will be retained for this domain. From the remaining domains, the model is recalculated and an outlier, if any, is again identified. The iterative process is repeated until no more outliers are found. In our investigations, we tried $\alpha = 0.05$.

5.4.2 MSE estimation

The Mean Square Error (MSE) is a useful concept for evaluating the gains of efficiency resulting from the use of the small area estimate $\hat{\theta}_i^{SAE}$ over the direct estimate $\hat{\theta}_i$:

$$MSE(\hat{\theta}_i^{SAE}) = E_{mp}(\hat{\theta}_i^{SAE} - \theta_i)^2.$$

The MSE is unknown but can be estimated. The MSE of the composite and synthetic estimators are calculated separately. The estimated MSE of the area level estimators depends on the procedure used for estimating the parameters (e.g., REML, ADM…). For the REML, the MSE of the small area estimate $\hat{\theta}_i^{SAE}$ can be calculated using:

$$mse(\hat{\theta}_i^{SAE}) = \begin{cases} g_{1i} + g_{2i} + 2g_{3i} & if\ i = 1, \dots, m \\ \boldsymbol{z}_i^T var(\hat{\boldsymbol{\beta}})\boldsymbol{z}_i + b_i^2\hat{\sigma}_v^2 & if\ i = m+1, \dots, M \end{cases}$$

where

$g_{1i} = \hat{\gamma}_i\hat{\tilde{\psi}}_i$,

$g_{2i} = (1-\hat{\gamma}_i)^2\boldsymbol{z}_i{}^T\left[\sum_{i=1}^m \boldsymbol{z}_i\,\boldsymbol{z}_i{}^T/(b_i^2\hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i)\right]^{-1}\boldsymbol{z}_i$,

$g_{3i} = (b_i^2\hat{\tilde{\psi}}_i)^2/(b_i^2\hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i)^3\ var(\hat{\sigma}_v^2)$,

$var(\hat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^m \frac{\boldsymbol{z}_i\boldsymbol{z}_i{}^T}{b_i^2\hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i}\right]^{-1}$ and $var(\hat{\sigma}_v^2)$ is the estimated asymptotic variance of $\hat{\sigma}_v^2$. The reference formulas are provided in Rao and Molina (2015, Chapter 6).

5.4.3 Coefficient of Variation

For evaluating the quality of estimates, coefficient of variation (CV) of the SAE estimates is calculated using the estimated mean square errors. The CV values are obtained by dividing the square root of the estimated mean square errors (MSE) of SAE estimates by the estimates:

$$CV_i = \frac{\sqrt{mse(\hat{\theta}_i^{SAE})}}{\hat{\theta}_i^{SAE}}.$$

6. Results

For the VTS data, the performance of the models varies by spending categories. For some spending categories, achieving a proper model fit is a challenge and the models need to be adjusted iteratively.

In general, for the smallest areas, the estimates are driven mostly by the predictions from the model. However, for the largest areas, the estimates tend to be close to the direct estimates. The following graph shows the CV of SAE and direct estimates for accommodation spending in Q2 (April-June) 2019, when domain are ordered by the estimated direct CV. The CV of small area estimates are smaller than direct estimates but the efficiency gain varies by domain. For a few domains, the EBLUP CV is larger than the direct CV. However, the estimated direct CV for these areas could be unreliable, as sample size of these domain is very small.

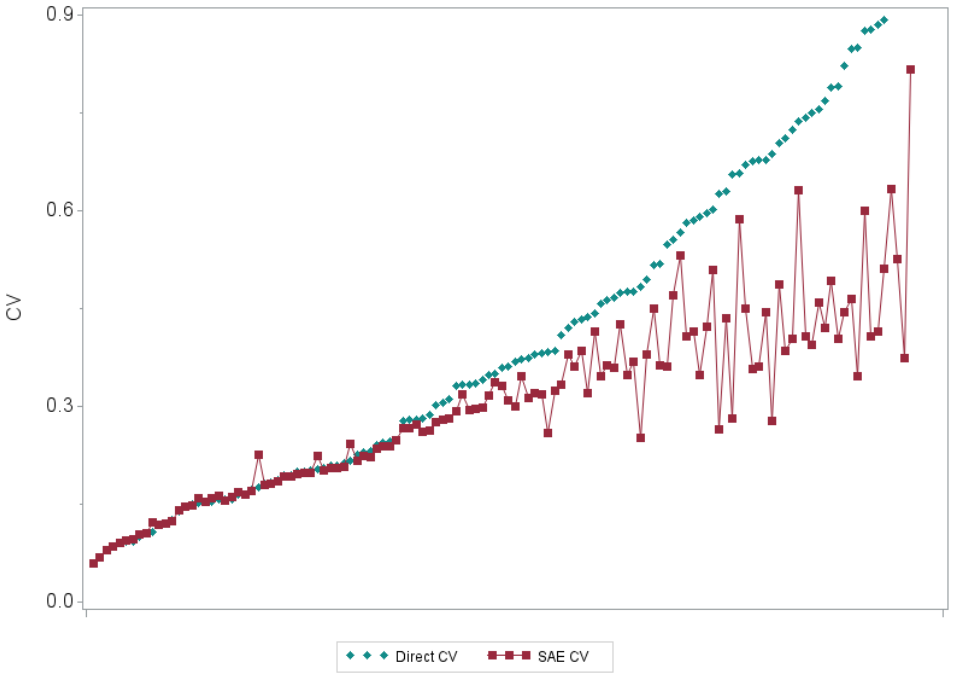

Figure 1: CV of SAE and direct estimates, accommodation spending Q2 2019

The small area estimates of the VTS are in general more efficient in areas with the smallest sample size. The efficiency is more substantial when γ values are close to zero but risk of bias due to model misspecification is larger in such cases. Figure 2 presents the relative

percent difference between SAE CV and direct CV or CV improvement (%) for accommodation spending in Q2 2019, when domains are ordered by domain sample size.

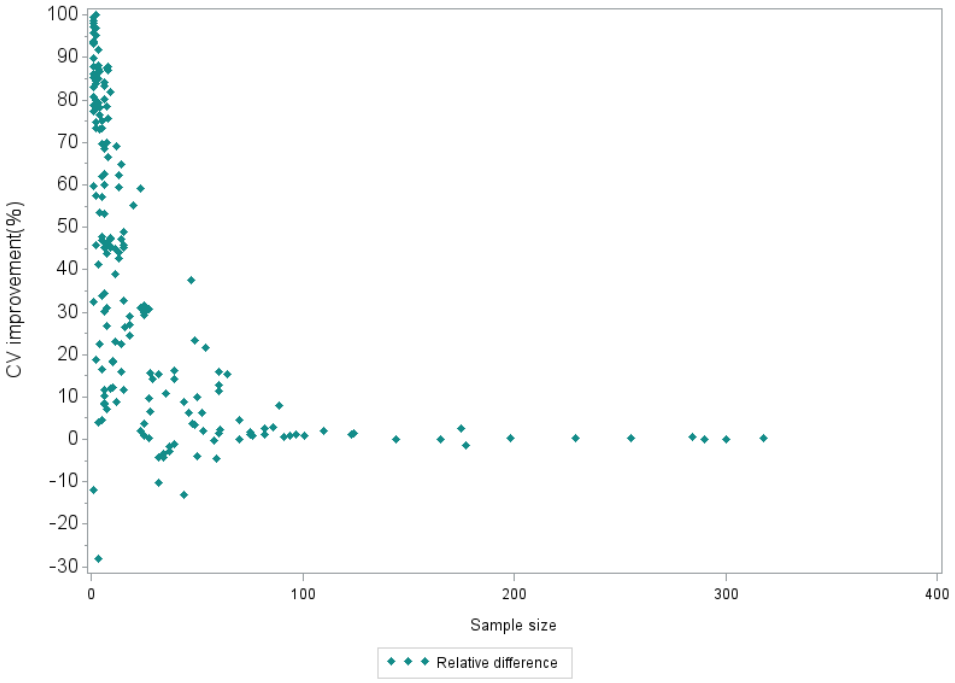

Figure 2: CV improvement (%) by domain sample size, accommodation spending Q2 2019

Aside from improvement in the efficiency, the number of available domain estimates increased. For domains with no sampled unit in the VTS, the small area estimates are synthetic estimates. For example, for the period of Q2 2019 the number of domain-level estimates increased by 31% in total. As the sample size increases, the relative percent difference between SAE CV and direct CV decreases.

7. Concluding Remarks

In general, the SAE methodology improves the quality of sub-provincial estimates. The comparison between direct and SAE estimates yielded varying results for different spending categories. Careful validation of assumptions and diagnostics plots are needed to avoid potential model misspecification. The basic models may need to be tailored to fit the data. For the VTS data, we had to use a modification of the FH model, piecewise area-level model.

Acknowledgements

The authors would like to thank Jean-François Beaumont, Cynthia Bocci and François Gagnon for their comments and suggestions that helped improve the content of this document.

References

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

Dick, P. (1995). Modelling Net Undercoverage in the 1991 Canadian Census, *Survey Methodology*, *21*(1), 45-54.

Estevao, V., You, Y., Hidiroglou, M., Beaumont, J.-F., and Rubin-Bleuer, S. (2019). Small Area Estimation-Area Level Model with EBLUP Estimation- Methodology Specifications. Statistics Canada document.

Fay, R.E. and Herriot, R.A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, *74*(366a), 269-277.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, *47*(260), 663-685.

Hidiroglou, M., Beaumont, J.-F., and Yung, W. (2019). Development of a small area estimation system at Statistics Canada. *Survey Methodology, 45*(1), 101-126.

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

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.

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

Rivest, L. P., and Belmonte, E. (2000). A Conditional Mean Square Error of Small Area Estimates. *Survey Methodology*, 26, 67-78.

Statistics Canada (2019). Small Area Estimation for Visitor Travel Survey. Available from: https://www.statcan.gc.ca/eng/statistical-programs/document/5261_D3_V2

Yung, W., (1997). Variance estimation for public use files under confidentiality constraints. Proceedings of the Survey Research Methods Section, American Statistical Association, 434-439.