Small Area Estimation for Business Surveys at Statistics Canada

Wesley Yung, Mike Hidiroglou and Victor Estevao

1.0 Introduction

Today's data users are becoming more and more demanding; they want more data, data at more detailed levels and they want them quicker. This demand is coming from many different sources such as government departments, policy makers and researchers. In the spring 2014 report of the Auditor General of Canada, it is recommended that 'Statistics Canada should assess the feasibility of more fully addressing user needs for data from small areas and subpopulations' (Office of the Auditor General of Canada, 2014). One obvious way to address this recommendation is to simply increase sample sizes. However, this is not a viable option given the current fiscal climate and the move to reducing respondent burden. A more innovative option is to combine administrative data with survey data through the use of small area estimation techniques. To this end, Statistics Canada has developed a SAS based small area estimation prototype to facilitate the production of small area estimates.

While small area estimation techniques have been used frequently with data from social surveys, their use in business surveys has been less common. When one thinks of small area estimation, one usually thinks of a small geographic domain such as a city or county. However, in business surveys, one is typically more interested in estimates for detailed industry domains than geographic domains. Business statistics programs also typically have access to very rich administrative data sources such as taxation data, which could lead to very high quality small area estimates. Thus, it is natural to consider small area estimation for business surveys.

In this paper, we introduce Statistics Canada's small area estimation prototype and present some results of its use with the survey on Research and Development in Canadian Industry (RDCI). The prototype and its functionalities are introduced in section 2. Details of the RDCI survey and small area estimation results obtained from the prototype are presented in section 3. Finally, some closing remarks are given in section 4.

2.0 Statistics Canada's Small Area Estimation Prototype

The small area prototype is a collection of SAS programs that produce estimates for either a unit-level or an area-level small area model. The programs consist of SAS macros and SAS IML modules and are designed to run under SAS 9.2 or 9.3 on a Windows-based platform. In the business survey context of estimation for industry domains, an area-level model could be used when auxiliary data are available at the industry level, while a unit-level model can be used when the auxiliary data are available for each business in the population of the industry.

Area-level models relate small area means to area-specific auxiliary data. They are viable when unit-level information is not available. More precisely, we define the sampling model as

$$\bar{y}_i = \theta_i + e_i, \;\; i = 1, \dots, m$$

where $\bar{y}_i$ is the estimated mean for small area *i*, *m* is the number of small areas, θ_i is the true mean for small area *i* and the e_i's are the corresponding sampling errors with $E_p(e_i|\theta_i) = 0$ and $V_p(e_i|\theta_i) = \psi_i$

(assumed known). The subscript *p* denotes that the expectation and variance are with respect to the sampling design. The linking model is given as

$$\theta_i = \boldsymbol{z}_i^T \boldsymbol{\beta} + \nu_i, \;\; i = 1, \dots, m$$

where $\boldsymbol{z}_i = (z_{i1}, \dots, z_{ip})^T$ is a vector of auxiliary data for small area *i*, $\boldsymbol{\beta}$ is a *px1* vector of regression parameters and the ν_i's are area specific random effects and are assumed to be independently and identically distributed (iid) with $E_\xi(\nu_i) = 0$ and $V_\xi(\nu_i) = \sigma_\nu^2 \geq 0$. The subscript ξ denotes that the expectation and variance are with respect to the model. Combining the sampling and linking models gives the area-level small area model

$$\bar{y}_i = \boldsymbol{z}_i^T \boldsymbol{\beta} + \nu_i + e_i, \;\; i = 1, \dots, m.$$

Note that it is assumed that the ν_i's and e_i's are independent.

In the small area prototype, two methods of estimation are available for area-level models: Empirical Best Linear Unbiased Predication (EBLUP) and Hierarchical Bayes (HB). For the EBLUP approach, the prototype offers four methods to estimate the variance components: adjusted density maximization (ADM) (Li and Lahiri, 2010), restricted maximum likelihood (Rao, 2003), Fay-Herriot (Fay and Herriot, 1979) and Wang-Fuller (Wang and Fuller, 2003).

For the HB approach, the prototype uses Monte Carlo Markov Chains (MCMC) with Gibbs sampling to estimate the model parameters. It offers three different linking models to allow for the most appropriate model:

- Matched Fay-Herriot model: $\theta_i = \boldsymbol{z}_i^T \boldsymbol{\beta} + \nu_i$
- Unmatched log-linear model: $\log(\theta_i) = \boldsymbol{z}_i^T \boldsymbol{\beta} + \nu_i$
- Unmatched log census undercount model: $\log\left(\frac{\theta_i}{\theta_i + C_i}\right) = \boldsymbol{z}_i^T \boldsymbol{\beta} + \nu_i$.

For more details on the models offered, see Estevao et al. (2014).

Unit-level models relate the business' values for the variable of interest to business-specific auxiliary data. These models require the availability of auxiliary data at the unit level for all units in the population. The prototype produces small area estimates through the use of a regression model with random area effects nested within the areas (a nested error model). We assume that unit-specific auxiliary data $\boldsymbol{x}_{ij} = (x_{ij1}, \dots, x_{ijp})^T$ are available for each unit *j* in small area *i* for *p* co-variates. The variable of interest, y_{ij}, is assumed to be related to $\boldsymbol{x}_{ij}$ through a linear regression model

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \nu_i + e_{ij}, \;\; i = 1, \dots, m, j = 1, \dots, N_i$$

where N_i is the population size of area *i* and the area specific effects, ν_i, are assumed to be iid with $E_\xi(\nu_i) = 0$ and $V_\xi(\nu_i) = \sigma_\nu^2 \geq 0$. The error terms e_{ij} are iid random variables with $E_\xi(e_{ij}) = 0$ and $V_\xi(e_{ij}) = \sigma_e^2$ and they are assumed to be independent of the ν_i's. Finally, it is often assumed that the ν_i's and the e_{ij}'s are normally distributed.

Two methods of estimation are available for the unit-level model: Empirical Best Linear Unbiased Prediction (EBLUP) and Pseudo-EBLUP that allows the use of survey weights. For a more detailed account of the methods mentioned above, the reader is encouraged to consult Rao (2003).

In addition to producing small area estimates, the prototype estimates a model-based Mean Squared Error (MSE) using the Prasad-Rao approach. Finally, the prototype includes numerous plots to verify the validity of the small area model and to evaluate the resulting small area estimates. Model diagnostics include residual plots, Q-Q plots of the residuals, influence measures of each small area and the Shapiro-Wilk test of standardized residuals. Evaluation plots include comparison of the estimates and MSEs with direct estimates and their corresponding variances.

3.0 Research and Development in Canadian Industry

Statistics Canada's RDCI survey collects information on expenditures on research and development (R&D) and personnel engaged in R&D activities. It is an annual survey of approximately 2,100 businesses in Canada. It uses a stratified Bernoulli design with the first level of stratification being 55 groups defined by the North America Industry Classification System (NAICS). Within each of these groups, units are further stratified into take-all, take-some and take-none strata based on their past expenses on R&D. Units in the take-all stratum are selected with certainty, while units in the take-some stratum are sampled with probabilities less than one. Units in the take-none stratum are not eligible for sampling. The sample design is optimized to produce reliable estimates for the 55 NAICS groups.

While the sample size of 2,100 units is adequate to produce estimates for the 55 NAICS groups, Statistics Canada's System of National Accounts (SNA) requires estimates for 212 detailed domains, which can not be reliably estimated with the current sample size. In reference year 2014, the RDCI sample size was increased to approximately 8,000 units, but this data is not yet available. We illustrate the small area prototype using data from reference year 2012. In addition to the frame information from Statistics Canada's Business Register, the RDCI enjoys the availability of administrative data from the Canada Revenue Agency (CRA) which administers a tax credit program for Scientific Research and Experimental Development Expenditures. Businesses engaged in R&D in Canada can apply for tax credits through CRA and the information collected is shared with Statistics Canada. One of the variables collected by CRA is the Capital Intramural Expenses (CIE), which is also collected by the survey. Note that the CRA data do not cover the entire population. Therefore, the small area prototype was used to combine the two sources of data to produce small area estimates.

Given that these two variables are available for each responding survey unit, the first model attempted was the unit-level model (see Rubin-Bleuer *et al.*, 2014). However, when looking at the two CIE values, there were many outliers that could have affected the performance of the small area model. Thus, it was decided to fit an area-level model to the data.

Small area estimates for the total CIE were produced for the take-some strata only for the SNA domains. In total, 654 sample respondents were available for the small area modeling which produced estimates for 188 of the 212 domains of interest. The corresponding variance estimates were smoothed, as is commonly done, to improve the small area estimates. Figure 3.1 presents a plot of the standardized residuals versus the auxiliary variable. It indicates that the residuals appear to be free of any patterns and are centered on zero (the red line). The blue line on the plot is a spline fit used to indicate the presence of discernible patterns in the data. It shows that there may be a couple of influential points,

but no obvious patterns. Two points stand out, which might be indicative of non-constant variance and will be further investigated with other diagnostic plots.

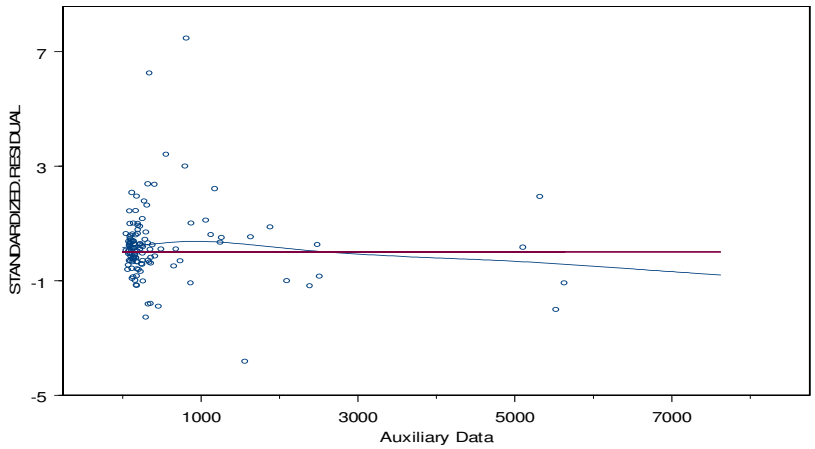

Figure 3.1 Plot of Standardized Residuals

The prototype also produces a diagnostic plot of influence measures, which is presented in figure 3.2. In this plot, two areas appear to have more influence than the remaining ones, but their influence measures are less than one so they are deemed to be not significant. These two areas correspond to the two large values of the standardized residual identified in figure 3.1 and, while influential, they do not appear to be outliers. An analysis of the Q-Q plot of the standardized residuals indicates that the assumption of normality does not appear to be violated.

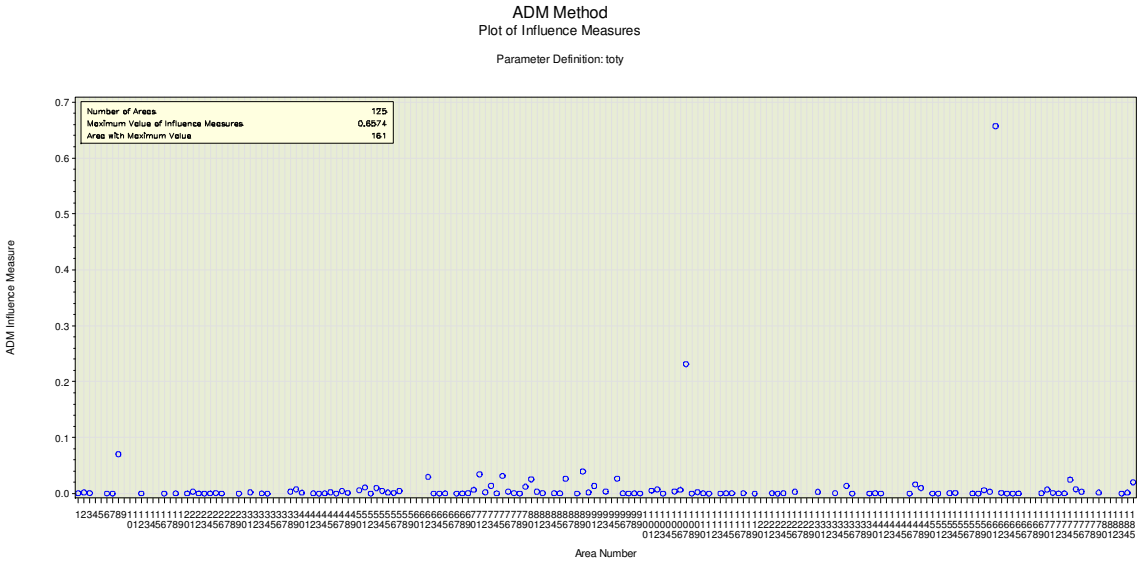

Figure 3.2 Plot of Influence Measures

The small area estimates for 188 domains of interest are given in figure 3.3, along with the direct survey estimates, as a function of the sample size in the small area. Note that estimates are produced for only 125 domains because in the remaining ones, the population or the sample size is zero.

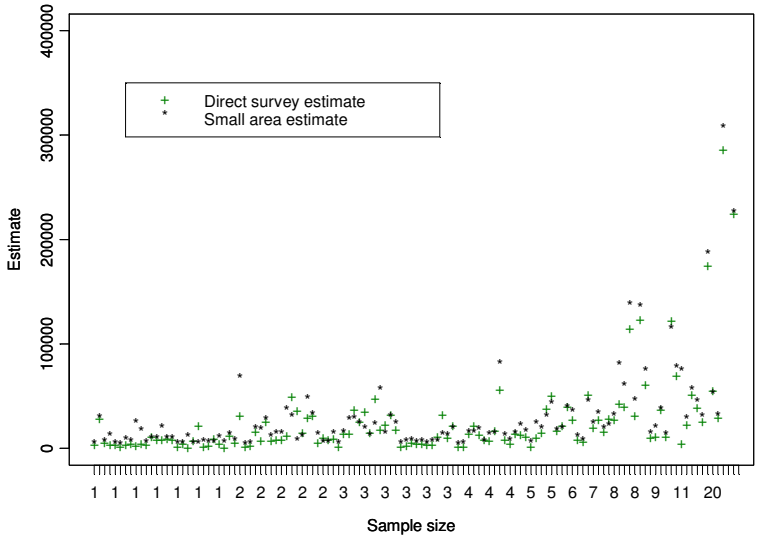

Figure 3.3 Plot of Small Area and Survey Estimates

As one can see, the small area and survey estimates are not systematically different and for most of the areas, the two are similar. Turning to quality indicators, figure 3.4 presents the square root of the estimated MSE of the small area estimates and the square root of the estimated survey variances, plotted against the sample size of the small area.

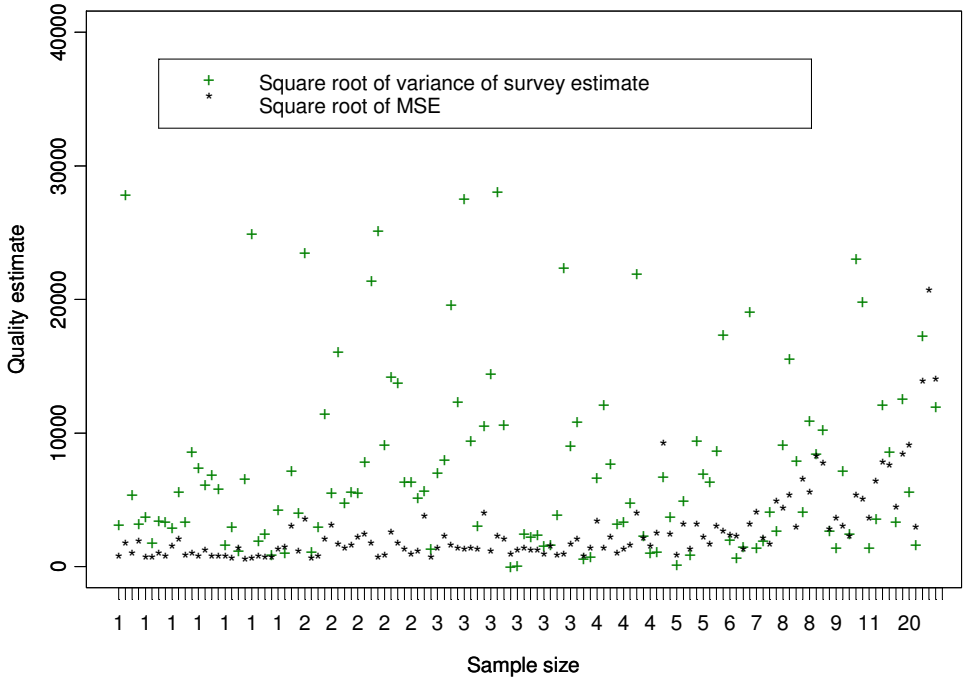

Figure 3.4 Plot of Estimates of Quality

As one can see, the small area estimates are more precise than the survey estimates when the sample sizes are small. As the sample sizes increase the precision of the two estimates becomes very similar, as expected

4.0 Summary

Based on the results obtained so far, small area estimation appears to be a viable option for producing estimates for domains that do not correspond to survey strata in business surveys. These domains can be of a geographical nature or based on detailed industry classifications. Statistics Canada's small area estimation prototype is a flexible system that offers both area-level and unit-level models, and several approaches to estimate the required small area model parameters. It also offers useful diagnostics to validate the small area models.

At this point in time, Statistics Canada is only investigating small area estimation methods and is not quite prepared to use the estimates for publication purposes. However, the plan is to use the RDCI small area estimates for reference year 2014 as experimental estimates for internal use by survey analysts and the SNA. These small area estimates will be produced at the same time as the survey estimates.

Acknowledgment: The authors would like to thank Jean-Francois Beaumont and Cynthia Bocci of Statistics Canada for sharing their results on the RDCI data.

References

Estevao, V. M., Hidiroglou, M.A., and You, Y. (2014). *Small Area Estimation - Area-level Model with EBLUP Estimation - Description of Function Parameters and User Guide.* Statistics Canada document.

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

Li, H., and Lahiri P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. Journal of Multivariate Analysis 101, 882-892.

Office of the Auditor General of Canada. (2014) Spring report. Minister of Public Works and Government Services Canada. Retrieved from: http://www.oag-bvg.gc.ca/internet/docs/parl_oag_201405_08_e.pdf.

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

Rubin-Bleuer, S., Julien, P.O., Pelletier, E. (2014). Feasibility Study on Small Area Estimation for RDCI, Internal Statistics Canada report.

Wang, J. and Fuller, W. A. (2003).The Mean Squared Error of Small Area Predictors Constructed With Estimated Area Variances American Statistical Association Journal of the American Statistical Association September 2003, Vol. 98, No. 463, Theory and Methods, 716-723.