Statistics Canada Statistique Canada

Canada

Statistics Canada
www.statcan.gc.ca

Small Area Estimation for Business Surveys at Statistics Canada

Wesley Yung, Mike Hidiroglou and Victor Estevao

Statistics Canada

European Establishment Statistics Workshop

Outline

- Introduction
- SAE Prototype
- Area Level Models
- Unit Level Models
- Illustration - RDCI
- Summary

Introduction

- Motivation for Small Area Estimation
  - Data users want more data, more details and want them now
  - Auditor General of Canada

    Statistics Canada should assess the feasibility of more fully addressing user needs for data from small areas and subpopulations
- Current environment has led to a need to find new ways
- A possible solution - Small Area Estimation (SAE)

Introduction

- SAE is not new at Statistics Canada
  - Used in for Census Undercoverage since 1991
- Recognizing the potential of SAE, Statistics Canada has developed an SAE prototype
- Business surveys have different characteristics
  - Interest doesn’t necessarily lie in geographical small areas
  - Interest typically lies in detailed industry domains (perhaps crossed with geography)
  - High quality auxiliary data is usually available
- This talk will present the prototype and illustrate its use with the survey of Research and Development in Canadian Industry (RDCI)

SAE Prototype

- Developed primarily by Mike Hidiroglou and Victor Estevao
- Consists of a series of SAS macros and IML modules
- Runs under SAS 9.2 or 9.3 in a Windows environment
- Is available free of charge but with limited support
- Handles both area level and unit level small area models

Area Level Models

- Area level models relate small area means or totals to area specific auxiliary data through models such as

$$\bar{y}_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i + e_i \qquad\qquad i = 1, \ldots, m$$

  where $\mathbf{z}_i$ is a vector of small area auxiliary data, $\boldsymbol{\beta}$ is a vector of regression parameters, v_i are the small area random effects and e_i are the sampling errors
- Certain assumptions on the model are made (Rao 2003)
- The prototype offers two methods of estimating the model parameters
  - Empirical Best Linear Unbiased Prediction (EBLUP)
  - Hierarchical Bayes (HB)

Area Level - EBLUP Estimation

- The small area model is a general linear mixed model
  - Can appeal to results from classical statistics to define the Best Linear Unbiased Predictor (BLUP) - assumes σ_v^2 known
- EBLUP is obtained by replacing σ_v^2 with an estimate $\hat{\sigma}_v^2$
- Four methods of estimating σ_v^2 are offered
  - Fay-Herriot method - a method of moments estimator
  - Restricted Maximum Likelihood (REML) - requires normality of v_i's
  - Wang-Fuller
  - Adjusted Maximum Likelihood

Area Level - HB Estimation

- The prototype has implemented the HB method using the MCMC method called Gibbs Sampling
- HB methods assume the Normality of the e_i's and v_i's in the area level model
- Under Normality and with σ_v^2 is known, the required conditional distributions can be derived
- Gibbs sampling is used to generate observations from these conditional distributions
- The SA estimates are calculated based on these observations
- For more details, see Rao (2003)

Area Level - HB Estimation

- The simple model assumes $\theta_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i$, which may not appropriate, so the prototype offers two additional linking models
  - Unmatched log-linear model $log(\theta_i) = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i$
  - Unmatched log census undercount model $log\left(\frac{\theta_i}{\theta_i + c_i}\right) = \mathbf{z}_i^T \boldsymbol{\beta} + b_i v_i$ where c_i is the number of units in small area i
- For more details on these conditional distributions, see Estevao et al. (2014)

Unit Level Models

- Unit level models relate a business' values to business specific auxiliary data
  - Auxiliary data must be available at the unit level for all units in the population
- For unit level models, the prototype uses a nested error model

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij} \qquad i = 1, \ldots, m, \ j = 1, \ldots, N_i$$

- This model is a special case of the general linear mixed model with block diagonal covariance structure
  - EBLUP methods for area level model are applicable
- Pseudo-EBLUP (uses survey weights) are also available in prototype
  - For details, see Rao (2003)

Additional Output

- Estimates of the MSE of small area estimates
  - Prasad-Rao approach for EBLUP
  - Gibbs sampler for HB
- Diagnostic plots
  - Residual plots
  - Q-Q plots
  - Influential measures

Illustration - Research and Development in Canadian Industry (RDCI)

- Collects information on R&D and personnel engaged in R&D
- Annual survey with $n \approx 2,100$
- Stratified Bernoulli design
- First level of stratification is 55 NAICS group
- Units further stratified by previous R&D expenses
- Sample design is optimized to produce estimates for the 55 groups
- System of National Accounts (SNA) requires estimates for 212 detailed domains

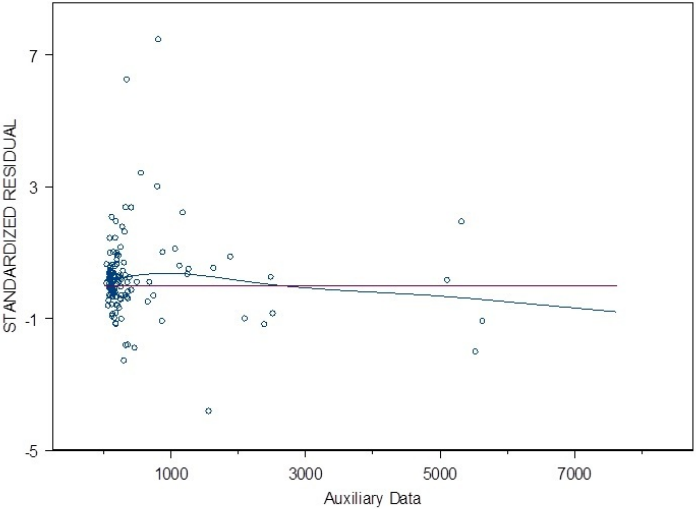
Illustration - RDCI

- Sample of 2,100 can not support this
- Auxiliary data are available from taxation agency
    - Tax credit for Scientific Research and Experimental Development Expenditures
    - Current Intramural Expenses (CIE) available on both survey and auxiliary data source
- SAE prototype used to produce estimates for 212 SNA domains
- The sample covered only 125 of these 212 domains

Illustration - RDCI

- Given availability of unit specific auxiliary data, unit level model was initially tried
  - Outliers severely affected the fit of the model
- Area level model was then used
- EBLUP estimation was used with variance components being estimated via ADM
- Diagnostic plots did not show any serious issues with model

Diagnostic Plots



Diagnostic Plots

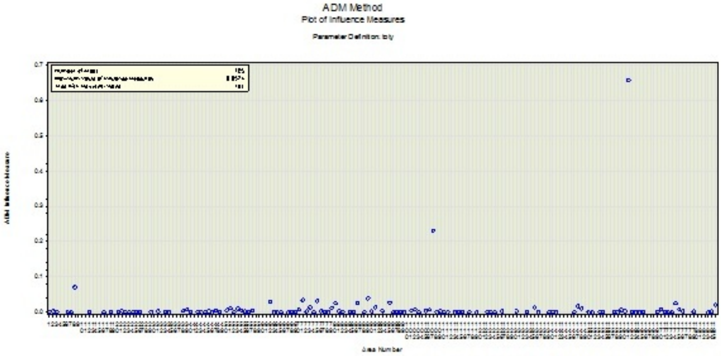
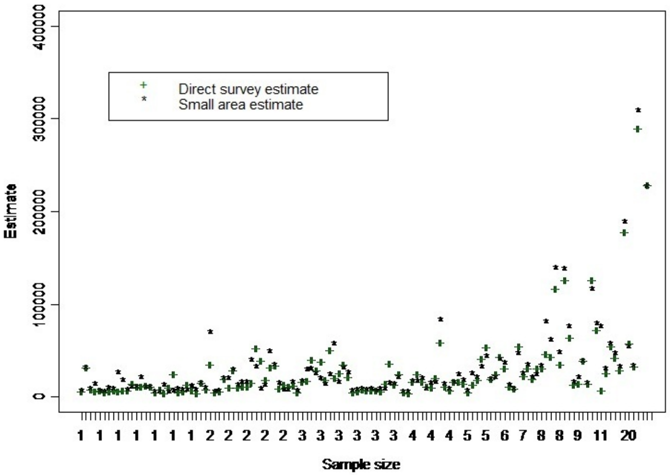


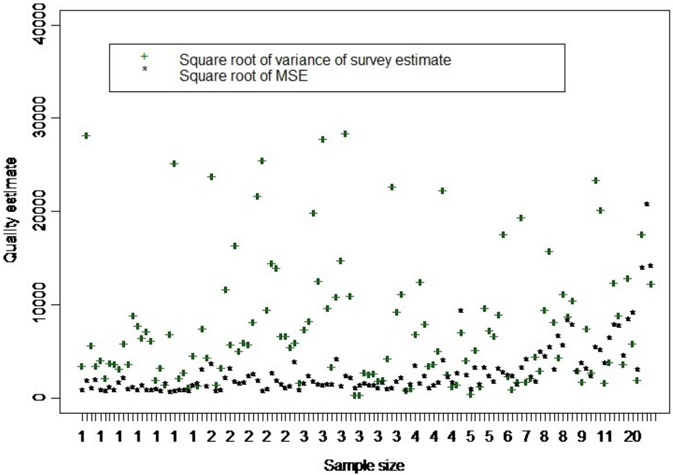
Illustration - RDCI

- SAE was able to produce estimates for 188 of the 212 domains
- In general, the direct survey and SAE estimates were similar
- Turning to MSE/Variance comparisons, the SAE performs much better when the sample sizes are small
- As sample sizes increase the two estimators behave similiarly

Direct vs Small Area Estimates



CV vs MSE


+ Square root of variance of survey estimate
* Square root of MSE
Quality estimate
0
10000
20000
30000
40000
Sample size
1 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 3 4 4 4 5 5 6 7 8 8 9 11 20

Summary

- Based on the RDCI example, SAE seems to be a viable option for business surveys
- Statistics Canada's prototype is a flexible SAE tool which offers users several options
- Statistics Canada is not yet prepared to publish SAE estimates but RDCI will produce SAE estimates for internal use for RY2014