

NEIGHBOURHOOD FACTORS AND CHILDREN: SMALL AREA STATISTICS

Longyang Wu, Xianlin Ma and Xu Wang¹

ABSTRACT

In this study we compare the design-based Horvitz-Thompson estimator to the model-based EBLUP estimator for estimating the 24 Census Metropolitan Area (CMA) means over a number of interested characteristics including measures of health conditions, injuries, and cognitive competence of children. The CMAs are treated as small areas since our sample sizes for the majority of them are smaller than fifteen. The area specific model at CMA level is adapted for model-based estimation due to the lack of complete information at the unit level. Our results show that the EBLUP estimator has smaller estimated mean square error compared to the design-based Horvitz-Thompson estimator.

KEY WORDS: Design-based estimator; EBLUP estimator; Imputation for missing data; Small area estimation.

RÉSUMÉ

Dans cette étude, nous comparons l'estimateur d'Horvitz-Thompson (basé sur le plan d'échantillonnage) avec l'estimateur EBLUP (basé sur un modèle) pour estimer les moyennes des 24 régions métropolitaines de recensement (RMR) pour un certain nombre de caractéristiques intéressantes comme les mesures de condition de santé, les blessures, et la compétence cognitive des enfants. Les RMR sont traitées comme de petits domaines étant donné que la taille d'échantillon est inférieure à 15 pour la plupart d'entre elles. Un modèle géographique spécifique au niveau des RMR est adapté pour l'estimation basée sur le modèle, dû au manque d'information au niveau de l'unité. Les résultats démontrent que l'estimateur EBLUP a une erreur quadratique moyenne estimée plus petite que l'estimateur Horvitz-Thompson.

MOTS CLÉS : Estimateur basé sur le plan d'échantillonnage; estimateur EBLUP; imputation pour données manquantes; estimation pour petits domaines.

1. INTRODUCTION

The data set used in this study is taken from the synthetic file released for cycle three of the National Longitudinal Survey of Children and Youth (NLSCY) conducted by Statistics Canada. The NLSCY is a long-term survey designed to measure child development and well-being. This survey selects representative samples of children from major metropolitan areas across Canada. The selected children are followed and monitored over time until their adulthood. In this study, we look at a particular data set for children aged 4, 5 or 6, living in one of 24 major metropolitan areas across Canada, with several measurements on neighbourhood factors and children's health conditions as well as cognitive competence.

Neighbourhood factors such as poverty and residential instability are likely associated with many health and educational problems related to children living in that particular area. Measures of health conditions and life opportunities of children are important to governments and researchers to have a better understanding of the ongoing life conditions of Canadian children and youth, and their developmental experiences.

¹ Longyang Wu, Xianlin Ma and Xu Wang, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1.
Contact Professor Changbao Wu (cbwu@uwaterloo.ca)

With the data we have, three variables are treated as responses: chronic health conditions, injury and cognitive competence. The NLSCY survey asked if the selected child had any health problems out of nine listed chronic health conditions such as Asthma, Allergies and so forth. Selected children are then grouped into three categories corresponding to 0 (no chronic conditions), 1 (only one) and 2 (2 or more). The survey also asked if children had serious injuries in the past 12 months. Injuries such as broken or fractured bones, burn or scald and so forth are considered as serious. Children with one or more serious injuries are labeled as 1 and 0 otherwise. The standard score measures children's cognitive competence by the revised Peabody Picture Vocabulary Test (PPVT-R). Scores are standardized to two-month age cohorts so that a score of 100 for a 5 year old is equivalent to a score of 100 for a 6 year old.

The data set also contains records over 13 other variables such as age, gender, number of siblings and so forth. Those variables are treated as auxiliary variables. In addition to this micro data file for which information is available at the individual level, we have access to a macro data file which provides summary statistics at the level of Census Metropolitan Area (CMA). The macro file contains five variables, namely the median share, the Gini coefficient, the proportion of persons below the poverty threshold of half the median, the coefficient of variation, and the median income. These variables reflect the socio-economic status of a CMA. It should be noted that variables appearing in the macro data file don't match any of those from the micro data file.

In this study, we focus on the estimation of population means at the CMA level using design-based and model-based approaches. Accurate and reliable estimation of such population means can provide information for policy makers and program officials to develop effective policies and strategies to help children and young people live healthy, active and rewarding lives. We investigate and compare results of the design-based Horvitz-Thompson (HT) estimator, the Generalized Regression Estimator (GREG), and the model-based EBLUP estimator for CMA means for each of the three aforementioned variables. The model-based EBLUP estimator for small areas is adopted because the direct design-based approach usually does not provide reliable results when the sample size is small. Generally speaking, the concept of small area does not necessarily mean geographically small. It refers to areas where the sample sizes are small. Under the current study, fifteen CMAs out of 24 have sample size of 15 or less, which makes the model-based small area estimation (SAE) approach appropriate. Therefore, we do not group the data further into bigger groups but use the natural CMAs as sub-populations in our analysis. Our major objective is to compare the direct design-based estimates and the model-based EBLUP estimates for the 24 CMAs.

In section 2 we outline the design-based estimator and the model-based small area estimator we use in our analysis. Section 3 presents the related results using data for the 24 CMAs. We conclude with a short discussion in Section 4.

2. METHODOLOGY

We have two working data files. The micro file contains basic information at the unit level for the three response variables as well as a number of auxiliary variables, with missing values occurring for some cases. The macro file contains summary information at the CMA level for five socio-economic status measures, but none of these variables is matched to any of the variables in the micro file. These features have to be taken into account when it comes to methodological considerations.

2.1 Handling Missing Values

There are about 1.5% of children injury data and 22% cognitive competence data which are missing. Two imputation methods were tried at the preliminary stage of our analysis for imputing missing values, namely, the nearest-neighbour imputation method and the regression imputation method. However, results from a preliminary regression analysis indicate that almost all auxiliary variables are not significant at 0.5% level for both variables of injury and cognitive competence. The missing mechanism seems not to depend on the auxiliary variables. We also compared our results when imputation is applied to those without imputation and they look very similar to each other. We have no further reason to believe that missing values depend on the response variable or the survey design. We hence consider the case here as missing completely at random (MCAR). For each of the three response variables, our analysis is based on data for which individuals with missing items are removed.

2.2 Design-based Estimators

Two commonly used design-based estimators are considered. The well-known Horvitz-Thompson (HT) estimator for the i th CMA mean is computed as

$$\hat{y}_{iHT} = \sum_{j \in s_i} w_{ij} y_{ij} / \sum_{j \in s_i} w_{ij},$$

where w_{ij} is the basic design weight (the inverse of the first order inclusion probability) and is available from the survey data, and s_i is the set of sampled individuals from the i th CMA. Since second order inclusion probabilities are not available, we use the following variance formula from sampling with replacement which is valid here if we assume that the sampling fraction is negligible:

$$v(\hat{Y}_{iHT}) \approx \frac{1}{\hat{N}_i^2} \frac{n_i}{n_i - 1} \sum_{j \in s_i} w_{ij}^2 (y_{ij} - \hat{Y}_{iHT})^2,$$

where $\hat{N}_i = \sum_{j \in s_i} w_{ij}$ is the estimated population total and n_i the sample size of the i th CMA.

With the availability of information on many auxiliary variables from the micro data file, it is tempted to consider the Generalized Regression Estimator (GREG) as an alternative design-based estimator:

$$\hat{y}_{iGR} = \hat{y}_{iHT} + \hat{B}_i^T (\bar{X}_i - \bar{X}_{iHT}),$$

where

$$\hat{B}_i = \left(\sum_{j \in s_i} w_{ij} x_{ij} x_{ij}^t \right)^{-1} \sum_{j \in s_i} w_{ij} x_{ij} y_{ij}.$$

The major difficulty in using the GREG estimator here is that the population means \bar{X}_i at the CMA level are not available from our data. One possible solution is to obtain a decent estimate of \bar{X}_i using a larger sample by combining adjacent CMAs together to form larger groups such as Western Canada, Ontario, Atlantic Canada and so forth, an idea parallel to the double sampling approach. We don't report our results under the GREG method, since variance estimation under this approach is very complicated, and the gain in efficiency compared to the HT estimator could be very marginal or even nil, given that the correlation between the response variables and the auxiliary variables are very weak under this particular study.

2.3 Small Area Estimation Using an Area Level Model

It is of great interest to consider model-based estimator under the framework of small area estimation, since our sample sizes for most CMAs are quite small. The lack of known population control totals (or means) for auxiliary variables in the micro data file and the availability of aggregated statistics at the CMA level from the macro data file make the use of an area level model a natural choice.

2.3.1 The Area Level Model

We adopt an area-specific model to estimate the area means for children chronic conditions, children injuries, and children's score of cognitive competence for each of the 24 CMAs. We utilize the area-specific (at CMA level) auxiliary data $z_i = (z_{i1}, \dots, z_{ip})^T$ available from the macro data file, and a specific function, $\theta_i = g(\bar{Y}_i)$, of the i th small area mean \bar{Y}_i , and link them with the following model:

$$\theta_i = z_i^T \beta + \nu_i, \quad i = 1, \dots, 24, \quad (1)$$

where the ν_i 's are random small area effects assumed to be iid with mean 0 and variance σ_ν^2 . We simply take $g(\bar{Y}_i) =$

\bar{Y}_i and use the direct estimator $\hat{\theta}_i = \hat{Y}_i$ of θ_i obtained through the HT estimator. We further assume that

$$\hat{\theta}_i = \theta_i + e_i, \quad i = 1, \dots, 24, \quad (2)$$

where e_i 's are sampling errors with mean 0 and variance ψ_i . These sampling variances can be estimated under the design-based framework and the estimated ψ_i are treated as known. Combing the sampling model (2) with the linking model (1), we have the following area-level model:

$$\hat{\theta}_i = z_i^T \beta + v_i + e_i \quad (3)$$

2.3.2 The EBLUP estimator

The empirical best linear unbiased predictor is used to estimate small area (CMA) means. If we assume that the errors v_i and e_i are independent and normally distributed, the Bayes posterior mean estimator is given by

$$E(\theta_i | \hat{\theta}_i, \beta, \sigma_v^2) = \gamma_i \hat{\theta}_i + (1 - \gamma_i) z_i^T \beta, \quad (4)$$

where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \psi_i)$ is the intra-class correlation. This Bayes estimator depends on β, σ_v^2 which are unknown in practice. However, for a given or estimated $\hat{\sigma}_v^2$, the generalized least squares estimator of β is given by

$$\hat{\beta} = [\sum_i z_i z_i^T / (\psi_i + \hat{\sigma}_v^2)]^{-1} [\sum_i z_i \hat{\theta}_i / (\psi_i + \hat{\sigma}_v^2)] \quad (5)$$

There are several methods available to obtain $\hat{\sigma}_v^2$. See, for instance, Rao (2003). We adopt the maximum likelihood method and use the score function to obtain the MLE iteratively as follows:

$$\sigma_v^{2(\alpha+1)} = \sigma_v^{2(\alpha)} + [(\sigma_v^{2(\alpha)})]^{-1} s(\hat{\beta}^{(\alpha)}, \sigma_v^{2(\alpha)}), \quad (6)$$

$$I(\sigma_v^2) = \frac{1}{2} \sum_i \frac{1}{(\sigma_v^2 + \psi_i)^2}. \quad (7)$$

The updating step is realized through (6), where the fisher information of $\hat{\sigma}_v^2$ is computed using (7). An R program is written to carry out the algorithm to obtain $\hat{\beta}$ and $\hat{\sigma}_v^2$ iteratively. After convergence, these two estimates are substituted into (4) to obtain the empirical Bayes estimator computed as

$$\hat{\theta}_i^{EB} = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) z_i^T \hat{\beta}. \quad (8)$$

The estimator given by (8) is also referred to as the Empirical Best Linear Unbiased Predictor (EBLUP) (Rao, 2003).

2.3.2 MSE of EBLUP

Under the normality assumptions for the error terms, it is shown in Rao (2003) that the mean square error (MSE) of $\hat{\theta}_i^{EB}$ can be approximated as the followings:

$$MSE(\hat{\theta}_i^{EB}) \approx g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2) \quad (9)$$

Where the three MSE components in this approximation are given by

$$\begin{aligned} g_{1i}(\hat{\sigma}_v^2) &= \hat{\sigma}_v^2 \psi_i / (\psi_i + \hat{\sigma}_v^2), \\ g_{2i}(\hat{\sigma}_v^2) &= (1 - \gamma_i)^2 z_i^T [\sum_i z_i z_i^T / (\psi_i + \hat{\sigma}_v^2)]^{-1} z_i, \\ g_{3i}(\hat{\sigma}_v^2) &= \psi_i^2 (\psi_i + \hat{\sigma}_v^2)^{-3} \bar{V}(\hat{\sigma}_v^2). \end{aligned}$$

The leading term $g_{1i}(\hat{\sigma}_v^2)$ is of order $O(1)$; the second term $g_{2i}(\hat{\sigma}_v^2)$ is of order $O(m^{-1})$, where m is the number of small areas in the analysis. In the last term, $\bar{V}(\hat{\sigma}_v^2)$ is the asymptotic variance of $\hat{\sigma}_v^2$.

3. RESULTS AND DISCUSSION

Figure 1 displays the estimated CMA means along with the estimated error for each one of the three response variable corresponding to children’s health conditions, injuries, and cognitive competence (i.e. voc. Score). It can be seen that the variability of these estimated means is smaller for the EBLUP estimator. The general trend is that the EBLUP estimates are centering around the “middle” of the HT estimates. This is probably due to the fact that the EBLUP estimator is the weighted average of the design-based HT estimator and the regression-synthetic estimator (the second component in (8)). This weighted average will shrink the extremely high or low CMA means towards the overall population mean. The computation of the weights involves the intro-class correlation γ_i for each CMA. This intro-class correlation measures the uncertainty in modeling the θ_i 's. If the model variance in (1) is small, then γ_i is small and more weight is attached to the synthetic estimator. On the other hand, more weight will be given to the direct estimator if the sampling variance ψ_i is relatively small. Our results also show that the EBLUP estimator has much smaller estimated variance when the CMA sample size is small. This point is further illustrated in Table 1, where the average of the estimated MSEs for the EBLUP estimator and HT estimator are compared for each of the three response variables. The average MSE for the EBLUP estimator is uniformly smaller for all cases.

Figure 1: Comparison of means and variances

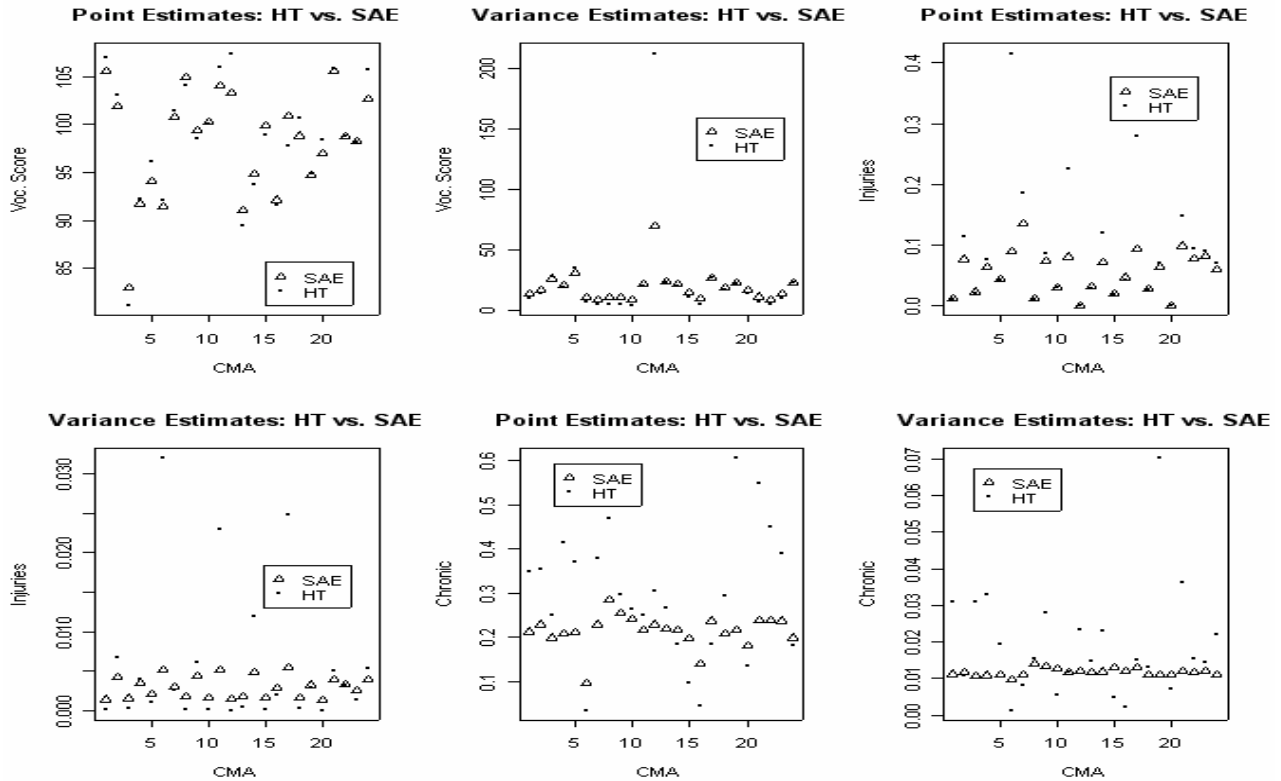


Table 1: Average Errors: SAE vs. HT

Method	Response Variables		
	Voc.Score	Injuries	Chronic
SAE	18.69	0.0030	0.012
HT	22.50	0.0055	0.019

ACKNOWLEDGEMENTS

This paper is written based on our presentation at the 2003 SSC annual meeting for the case study on neighbourhood factors and Children under the supervision of Professor Changbao Wu. We would like to express our thanks to Professor Ng for her help in the case study registration process; to Professors Welch and Chipman for their kind financial help; to the hosts at Dalhousie University for organizing a wonderful conference.

REFERENCES

- Lohr, Sharon L. (1999). *Sampling Design and Analysis*. Duxbury.
- Rao, J.N.K (2003). *Small Area Estimation*. New York. Wiley.