STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 53–58, DOI 10.21307/stattrans-2020-029
Received – 31.01.2020; accepted – 30.06.2020

Discussion of "Small Area Estimation: Its Evolution in Five Decades", by Malay Ghosh

J. N. K. Rao[1]

1. Introduction

It is my great pleasure to act as an invited discussant of this overview paper on small area estimation (SAE) by Malay Ghosh, based on his 28th Annual Morris Hansen Lecture held on October 30, 2019 in Washington, D.C. I was closely associated with the late Morris Hansen while we were both members of the Statistics Canada Methodology Advisory Committee for several years chaired by Hansen. I greatly benefited from his pioneering contributions to survey sampling theory and practice. Ghosh and I collaborated on a SAE review paper 26 years ago (Ghosh and Rao, 1994), which has received more than 1000 Google citations and partly stimulated much research on SAE over the past 25 years. The greatly increased demand for reliable small area statistics worldwide of course is the primary factor for the explosive growth in the SAE methodology. My joint paper with Ghosh stimulated me to write my 2003 Wiley book on SAE (Rao 2003). Because of the extensive developments in SAE after my 2003 book appeared, I wrote the second edition of my Wiley book in 2015 jointly with Isabel Molina (Rao and Molina 2015). Perhaps, my 2015 book is now obsolete given the rapid new developments in SAE theory and practice over the past 5 years!

Direct area-specific estimates are inadequate for SAE due to small domain or area sample sizes or even zero sample sizes in some small areas. It is therefore necessary to take advantage of the information in related areas through linking models to arrive at reliable model-dependent or indirect small area estimates. Hansen et al. (1983) demonstrated that model-dependent strategies can perform poorly for large samples even under small model misspecification, unlike design-based strategies leading to design-consistent estimators. On the other hand, Hansen et al. (1983) also note that the model-dependent strategies might enjoy substantial advantage in small samples if the model is appropriate and the sampling plan need not be probability based. The latter statement has implications to current focus on non-probability samples. Kalton (2018)

[1] School of Mathematics and Statistics, Carleton University, Canada. E-mail: jrao@math.carleton.ca. ORCID: https://orcid.org/ 0000-0003-1103-5500.

says "Opposition to using models has been overcome by the demand for small area estimates".

Ghosh provides a nice overview of methods for indirect estimation of small area means or totals over the past 50 years, starting with the use of synthetic estimation in the context of a radio listening survey (Hansen et al. 1953, pp. 483–486). In the early days, indirect estimates were based on simple implicit linking models (Rao and Molina, 2015, Chapter 3), but methods based on explicit linking models have taken over because of many advantages including the following: (a) model diagnostics to find suitable models can be implemented, (b) area-specific estimates of mean squared error (MSE) can be associated with each small area estimate, unlike the global measures of precision (averaged over small areas) often used with traditional synthetic estimates, and (c) "optimal" estimates of small area parameters under linear mixed and generalized linear mixed models can be obtained using empirical best unbiased prediction (EBLUP), empirical best (EB) or hierarchical Bayes (HB) methods. The HB method is currently popular because of its ability to handle complex models in an orderly manner and the availability of powerful computer programs to implement sophisticated HB methods. Ghosh has made significant contributions to the HB method for SAE. It is interesting to note that his first two papers on HB were jointly with his former students Partha Lahiri and Gauri Datta (Ghosh and Lahiri 1989 and Datta and Ghosh 1991). As we all know, both Lahiri and Datta have become leading researchers in SAE.

2. Basic area-level model

For simplicity, Ghosh focused his paper on the basic area level model (also called the Fay-Herriot model) in sections 5, 7 and 8 supplemented by a brief account of model based SAE under a basic unit level nested error linear regression model (also called the Battese-Harter-Fuller model) in Section 6. He presents the empirical best linear unbiased predictor (EBLUP) which avoids the normality assumption, using the moment estimator of the random effect variance A proposed by Prasad and Rao (1990). He also gives the estimator of the mean squared prediction error (MSPE) proposed by Prasad and Rao (PR), which is second-order unbiased for the MSPE, under normality assumption. He also mentions the work of Lahiri and Rao (1995), which proved the second-order unbiasedness of the PR MSE estimator without normality assumption on the random area effects in the model, provided the PR moment estimator of A is used. Fay and Herriot (1979) proposed a different moment estimator of A by solving two equations iteratively.

The moment estimators of A as well as the maximum likelihood (ML) and the residual ML (REML) estimators might produce zero estimates. In this case, the EBLUPs will give zero weight to the direct estimates in all areas, regardless of the efficiency of the direct estimator in each area. On the other hand, survey practitioners often prefer to give always a strictly positive weight to direct estimators because they are based on the area-specific unit level data without a model assumption. For this situation, Li and Lahiri (2010) proposed an adjusted ML (AML) estimator that delivers a strictly positive estimator of A . Molina et al. (2015) proposed modifications of the AML estimator that use the AML estimator only when the REML estimator is zero or when the data does not provide enough evidence against the hypothesis. Their simulation study suggested that the EBLUPs based on the modified estimators of A lead to smaller average MSE than the A AML-based EBLUPs when A is small relative to the variance of the direct estimator. They also proposed an MSE estimator that performed well in terms of average absolute relative bias even when A is small relative to the variance of the direct estimator.

In my books I emphasized the need for external evaluations by comparing the small area estimates to corresponding gold standard values, say from the recent census, in terms of absolute relative error (ARE) averaged over groups of areas, where ARE for a specific area is equal to |est. – truth|/truth. Ghosh mentions an external evaluation in the context of estimating median income of four-person families for the 50 states and the District of Columbia in USA. His Table 1 shows that the EBLUP leads to significant reduction in ARE averaged over the areas relative to the corresponding direct estimate obtained from the Current Population Survey (CPS). Hidiroglou et al. (2019) report the results of a recent external evaluation on Canadian data. Here Census Areas (CAs) are small areas, direct estimates are unemployment rates from the Canadian Labor Force Survey (LFS) and Employment Insurance (EI) beneficiary rate is the area level covariate, which is an excellent predictor of unemployment rate. Direct estimates from a much larger National Household Survey (NHS) were treated as gold standard or true values. The external evaluation showed that for the 28 smallest areas ARE for the LFS estimates is 33.9% compared to 14.7% for the EBLUP. Statistics Canada is now embarked on a very active SAE program and the demand for reliable small area estimates has greatly increased.

EBLUP-type model dependent estimates are often deemed suitable by National Statistical Agencies to produce official statistics, after careful external evaluations as mentioned above. However, those agencies often prefer estimators of design mean squared error (DMSE) of the EBLUP rather than its estimator of model-based MSPE, similar to estimators of DMSE of the direct estimator, conditional on the small area parameters , see Pfeffermann and Ben-Hur (2019). Exact design-unbiased estimator of EBLUP can be obtained but it is highly unstable due to small sample size in the area and also it can take negative values often when the sampling variance of the direct estimator is large relate to the model variance of the random area effect (Datta et al., 2011). Recent research attempts to remedy the difficulty with the design unbiased estimator. Rao et al. (2018) proposed a composite estimator of design MSE of EBLUP by taking a weighted combination of the design-unbiased MSE estimator and the model-based estimator of MSPE, using the same weights as those used in constructing the EBLUP as a weighted sum of the direct estimator and the synthetic estimator. It performed well in simulations in overcoming the instability associated with the design unbiased MSE estimator and reducing the probability of getting negative values. Pfeffermann and Ben-Hur (2019) proposed an alternative estimator of DMSE of EBLUP, based on a bootstrap method restricted to the distribution generated by the sampling design.

3. Some extensions

Ghosh mentions an extension of the basic FH model that allows different random effect variances for different small areas. In this case, he refers to the HB method of Tang et al. (2018) based on "global-local shrinkage priors", which can capture potential "sparsity" by assigning large probabilities to random area effects close to zero and at the same time identifying random area effects significantly different from zero. Ghosh mentions that such priors are particularly useful when the number of small areas is very large. I believe this extension is very useful and I expect to see further work on this topic.

Ghosh lists several important topics not covered in his review, including robust SAE, misspecification of linking models and estimation of complex area parameters such as poverty indicators. I will make few remarks on the latter topics.

An excellent review paper by Jiang and J. S. Rao (2020) covers robust SAE and model misspecification. They mention the work of Sinha and Rao (2009) on robust EBLUP (REBLUP) under unit level models that can provide protection against representative outliers in the unit errors and/or area effects. Dehnel and Wawrowski (2020) applied the REBLUP method to provide robust estimates of wages in small enterprises in Poland's districts. Jiang and J. S. Rao (2020) also mention their earlier work (Jiang et al. 2011) on misspecification of the linking model under the FH model.

Most of the past work on SAE focused on area means or totals under area level and unit level models. However, in recent years the estimation of complex small area parameters has received a lot of attention, such as small area poverty indicators that are extensively used for constructing poverty maps. We refer the reader to a review paper (Guadadarrama et al. 2014) and Rao and Molina (2015, Chapter 9) on estimating poverty indicators proposed by the World Bank: poverty rate, poverty gap and poverty severity. They studied empirical best or Bayes (EB) and HB methods and compared them to a method used by the World Bank, called ELL method.

There is also current interest in using estimates from big data or nonprobability samples as additional predictors or covariates in area level models. Rao (2020) mentions some recent applications of using big data as covariates.

4. Production of small area official statistics

Tzavidis et al. (2019) provide a framework for production of small area official statistics using model-dependent methods. Molina and Marhuenda (2015) developed an R package for SAE that was used in the book by Rao and Molina (2015).

REFERENCES

DEHNEL, G., WAWROWSKI, L., (2020). Robust estimation of wages in small enterprises: the application to Poland's districts, Statistics in Transition new series, 21, pp. 137–157.

GUADARRAMA, M., MOLINA, I., RAO, J. N. K., (2016). A comparison of small area estimation methods for poverty mapping, Statistics in Transition new series, 1, pp. 41–66.

GHOSH, M., LAHIRI, P., (1989). A hierarchical Bayes approach to small area estimation with auxiliary information, In: Proceedings of the Joint Indo- US Workshop on Bayesian Inference in Statistics and Econometrics.

HANSEN, M. H., MADOW, W. G., TEPPING, B. J., (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys, Journal of the American Statistical Association, 78, pp. 776–793.

HIDIROGLOU, M. A, BEAUMONT, J-F., YUNG, W., (2019). Development of a small area estimation system at Statistics Canada, Survey Methodology, 45, pp. 101–126.

JIANG, J., RAO, J. S., (2020). Robust small area estimation: An overview, Annual Reviews, 7, pp. 337–360.

KALTON, G., (2019). Developments in survey research over the past 60 years: A personal perspective, International Statistical Review, 87, pp. S10–S30.

LI, H., LAHIRI, P., (2010). An adjusted maximum likelihood method for solving small area estimation problems, Journal of Multivariate Analysis, 101, pp. 882–892.

MOLINA, I., MARHUENDA, Y., (2015). Sae: An R package for Small Area Estimation, The R Journal of Statistics, 7, pp. 81–98.

MOLINA, I., RAO, J. N. K., DATTA, G. S., (2015). Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random effects, Survey Methodology, 41, pp. 1–19.

PFEFFERMANN, D., BEN-HUR, D., (2018). Estimation of randomization mean squared error in small area estimation, International Statistical Review, 87, pp. S31–S49.

RAO, J. N. K., (2003). Small Area Estimation. Hoboken, NJ: Wiley.

RAO, J. N. K., (2020). On making valid inferences by integrating data from surveys and other sources, Sankhya, Series B (in press).

RAO, J. N. K., RUBIN-BLEUER, S., ESTEVAO, V. M., (2018). Measuring uncertainty associated with model-based small area estimators, Survey Methodology, 44, pp. 151–166.

SINHA, S. K., RAO, J. N. K., (2009). Robust small area estimation, Canadian Journal of Statistics, 37, pp. 381–399.

TZAVIDIS, N., ZHANG, L. C., LUNA, A., SCHMID, T., ROJAS-PERILLA, N., (2018). From start to finish: a framework for the production of small area official statistics, Journal of the Royal Statistical Society, Series A, 181, pp. 927–979.